Employing a comparative evaluation of Heuristic evaluations with end-users and usability experts as evaluators

MADELEINE MS SILVERBRATT

Uppsala University, madeleine.silverbratt@gmail.com

There is scarce research that implement a formal framework when evaluating usability evaluation methods such as heuristic evaluation. This paper aimed to explore and compare the results of a heuristic evaluation performed by end-users and a heuristic evaluation performed by experts. Both heuristic evaluations took place in the context of forestry industry where a mobile application developed to give harvest operators performance feedback was evaluated. A thorough literature review for research regarding evaluation of UEM was a crucial first step. The outcome of this produced an evaluation framework that included three criteria, Relevance, Frequency and Timeliness. These criteria were used to analyse the results from the heuristics evaluations performed by the two groups, using mixed methods. The quantitative analysis concluded that the evaluation performed by the end-users had a higher frequency and relevance value, and that the evaluation performed by the expert group had higher value for their solution rate in the timeliness criteria. Furthermore, the qualitative analysis held within the criteria timeliness concluded that the two groups identified different types of usability problems, confirming previous research performed on different types of heuristic evaluators.

Additional Keywords and Phrases: Usability Evaluation Method; Heuristic Evaluation; Mixed Methods; Relevance, Frequency; Timeliness

ACM Reference Format:

1 INTRODUCTION

This research study was conducted within the forestry industry in collaboration with the research institute Skogforsk and the forestry organisation Vida. In a study performed by Skogforsk to investigate the harvester operators decision-making, the results showed that the harvest machinery can make better calculated decisions regarding where to cut the trees compared to harvest operators [92]. Furthermore, to unlock the harvesters full potential, a support system for the harvester operating the machines is necessary [93]. To solve this issue, Vida developed an app where harvester operators (that work on commission with Vida) receive their feedback of what they have produced from their harvesting. The mobile application Vida WQP, was created one year ago, and new functions and alterations are continuously added [48]. It is important that this app is understandable and relevant from both Vida's and the harve+ter operators' perspectives since, the purpose of this system is to provide feedback

to their harvest operators (users). Addressing the issue of how to present feedback to the harvester operators is an interest for the forestry industry itself, therefore, both Vida and Skogforsk have an interest in evaluating the mobile application Vida WQP´s usability by conducting usability evaluation methods (UEMs) [75].

UEMs are a collection of methods that are a part of the User-Centred Design (UCD) field [76]. UCD is an iterative user-driven design process where all decisions and development consider the usability and user satisfaction of a system in all steps. According to ISO 9241-210:2010, a core step in UCD is that the iterative design process is based on the ongoing evaluation from users [4]. Their purpose is to answer how effective and efficient the system is, and how the users experience the interactive system [6]. Examples of usability evaluation methods are Concurrent Think-Aloud, where the participants vocalise their thoughts as they perform their given tasks and constructive interaction, where two or more participants work together with the tasks they were given [19,79]. Other UEMs that are common and utilize less resources are usability questionnaires such as User Experience Questionnaire (UEQ), System Usability Scale (SUS) and Heuristic Evaluation (HE) [48]. Usability questionnaires are mostly employed to measure the current usability state of a system, whereas HE will identify and provide solutions for usability problems rather than a specific score value [75].

Choosing the correct UEM to evaluate a systems usability is important and needs to utilize the right resources to generate the best result possible, meaning that the UEM should identify as many of the usability problem a system has [28.77]. Which makes it important to evaluate and compare UEMs to see which method is most adequate to use. HEs are appropriate to use in scenarios where the aim is to discover usability problems and their individual elements and discover how they impact the overall user experience (UX). UEMs such as HEs can be applied in two ways, either in a facilitated laboratory study or in case studies. According to Moumane et al., case studies allow the users to interact and perform tasks in a realistic manner [58]. Therefore, to fulfil this study´s aim of comparing the two evaluations the study assumed a case study approach [11.44.82], this approach ensured that enough detailed data could be collected to compare and evaluate the two evaluations. Therefore, two HEs evaluated Vida WQP´s usability, the first evaluation was conducted by Vida WQP´s users and the second evaluation was conducted by HCI master students, whom in the scope of this study are considered usability experts [10].

2 BACKGROUND AND THEORY

2.1 Vida´s feedback application: WQP

The forestry industry's clients have become more specific about the length and diameter combinations they want for their timber, so the logs can be used in the most efficient way possible. The specific orders means that the decisions that harvester operators make in the field affect the value of the products that are produced. To maximise their production, a harvester operator has a computer that calculates what logs are to be made, but the operators still must make decisions if a situation should require it. For example, if trees are bark beetle damaged which requires the harvest operators to make manual decisions.

Today's forestry machines are mainly run by computed decisions, but some decisions must be made by the harvest operator. The goal is to interfere as little as possible with the machine's choices, that is why a tool to help harvest operators accommodate what has been ordered and reduce the amount of inaccurate timber has been implemented.

Vida WQP is an in-house app developed by Vida with the aim to easier supply their harvest operators with feedback on what they have produced. Being able to return feedback is an important step to both the harvest

operator and Vida [92]. Vida's app has three different areas to distribute feedback to their harvest operators (A1). The first area is called Harvested and contains four functions where the operators can see what they have produced in their forestry machines. The second area shows what the harvest operator production is when it is measured at the sawmill. The third area stores a pdf document of measurements provision that the harvest operator should proceed from depending on what type of timber it is and what type of product it is supposed to be. The harvested and the measured area has either a green, yellow or red light next to each function. The light depends on if what the harvest operator has produced is within a specific value. This is meant to illustrate and give feedback to the harvest operators and what they have produced.

2.2 Heuristic Evaluation

HE is frequently applied in both software development and user interfaces to test their usability [66]. It has been used as a guidance for design decisions since they were introduced in 1994 by Nielsen [61]. In a study by Nielsen and Landauer they present a model of how the number of experts have a correlation to how many of the usability problems that will be identified [62]. Their prediction formula shows that 5 evaluators will uncover 75% of the usability problems [66], compared to using only one evaluator which according to the prediction formula uncover approximately 35%. Therefore, Nielsen argues that using a minimum of five evaluators increases the effectiveness of the HE method.

2.2.1 Defining HE Experts

HE is a cost-effective way to evaluate and test software products [50,51]. Nielsen argues that a successful outcome of the evaluation is heavily connected to the knowledge of the evaluators [32,63], hence this type of method requires experts as the evaluators [16,50,62,66]. In a systematic literature review Paz et al. concluded that different profiles of experts were used in HEs [66]. In their study they identified five fields, usability specialists, software professionals, domain experts, usability specialists combined with domain experts and double experts. Examples of domain experts are, people that have a profession within the software product that is evaluated, users that are connected to the software product evaluated and people that use the software product that is under evaluation [31,43,74,90]. Studies that used double experts defined them as people that are usability specialists as well as experts within the domain that is under evaluation [2]. Hassan et al. used different types of experts depending on which scope the evaluations were employed in [28]. They used software experts in a HE during the development of the software, and end-users as experts during the finishing process of the software product that was under evaluation.

The term expert is broadly applied in the context of HE [48]. There are no set of requirements that have to be fulfilled by the experts. Paz et al. identified evaluations that used experts from different types of and when Nielsen performed an empirical test of HE as an UEM he used 37 computer science students that had a lecture on evaluating interfaces through heuristics before the evaluation took place [63,66]. In a study performed by Gulliksen et al. they concluded that there is a relationship to what types of problem a usability specialist tend to identify and their background and experiences [24]. Furthermore, it is important that they have experience of building solutions of interaction design in terms of structure, content, concept and navigation. Botella et al. defined credentials that a usability expert should obtain or acquire to be considered a usability expert [10]. They argue that the experts used for HE should have deep knowledge in the field of HCI, the field of software engineering and how a system development life cycle in large projects works. Furthermore, their study concluded that the experts should have

experience in how to apply that knowledge in methods and analysis such as interviews, observations and surveys with users. Besides knowledge and experience in the field of usability Botella et al. states that they should have capacities in fields such as business, design or psychology and have experience in the field of design, both user interfaces and visual design guides.

Employing HE using novice evaluators instead of expert evaluators can enable organisations that lack the funds to conduct evaluations using experts [51]. The requirements of using experts partially originates from the complexity of the established heuristics [39,77]. Due to the complexity, novice evaluators have difficulties understanding and separating the heuristics from each other. De Lima Salgado & de Mattos Fortes states that there are studies conducted on HE that are using novice evaluators and that the heuristics given to them have been adapted [46]. The few studies that exist today have adapted the heuristics to specific profiles of inexperienced evaluators and can therefore not be generalised [71].

2.2.2 Choosing the appropriate set of heuristics

Bertini et al. [8] released heuristic guidelines that were appropriate to use in a mobile-computing context. Some of the modified heuristics were specifically adapted for mobile phones and their physical elements. However, these heuristics were based on mobile computing in 2006, and some of the heuristics are therefore not relevant in usability evaluation performed on smartphones [16]. In 2012 Inostroza et al. introduced a set of 12 heuristics to be used when evaluating touchscreen-based mobile devices [35]. Although these heuristics are more up to date, they are developed to evaluate the interface of the phone and not to evaluate mobile applications. Therefore, Pimento and Neto´s set of heuristics are appropriate to use [53]. They developed and validated 11 heuristics applicable for mobile applications, they are inspired by Nielsen´s set of heuristics but simplified and adjusted them for mobile interfaces and their changed environment (A3).

2.3 Evaluating UEMs

Wixon argues that to define the best UEM it is important to evaluate the UEMs by the appropriate criteria [85]. From previous literature he concluded that there were three premises all studies shared when evaluating UEMs. The premises were dismissed as inappropriate by Wixon since he means that criteria should take the practical work and developments of real products in commercial enterprises into account. Therefore, criteria that build on the amount of problems detected allow the UEMs to be evaluated in isolation to the case, and apply a quasi-scientific framework to resolve what UEM is the best are not appropriate to use according to Wixon [23]. Instead, he suggests that the correct criteria should be built upon business and engineering XXX which would be more in line with the values and limitations of the current practice, since a scientific approach would be incongruent with the context of the product development.

Wixon states that the credibility and timeliness of the integrated usability with the general development and process of the product development, these are criteria that should be considered when evaluating UEMs [85]. Credibility is considered a multi-dimensional concept, involving source of information and how trustworthy the information is regarding its trustworthiness, expertise, credentials and more [26]. In a study that assessed a computation model through isolated testing credibility was defined as something qualified to be trusted and believed in, which has been obtained through gathering and presenting evidence [57]. Timeliness is considered to be up-to-date and contain well-timed and relevant information [26], furthermore Rashtian & Gopalakrishnan defined timeliness as keeping a deadline and time-sensitive data [70].

In a world where usability work is applied and integrated in a business framework, a factor that should be in focus when evaluating UEMs is success. Success in this context means how effective the usability improvements deriving from the usability problems can be implemented in the product. Furthermore, Wixon argues that a case study is the most effective way to evaluate and compare UEMs since it does so in a practical manner. He argues that evaluating UEMs in a laboratory setup fails to include how practical realities of software and user interface development affect the implemented UEM [65].

To summarise, Wixon proposes three key notes when evaluating a UEM [85]. It is necessary to adopt relevant business and engineering criteria, such as credibility, timeliness and success. The UEM needs to be evaluated in a real setting applied to real products. The evaluation should follow the approach of a case study rather than an experimental approach.

Van den Haak et al. did a methodological comparison with three UEMS, two different think-aloud methods and constructive interaction [79]. Their study was focused on identifying and describing the UEMs as tools for usability testing rather than uncovering cognitive approaches. They evaluated the UEMs from four different perspectives. The first perspective aimed to identify the amount and types of usability problems detected by each UEM. The second perspective aimed to identify the relevance of the usability problem detected by each UEM. The third perspective aimed to identify how the three UEMs differed in terms of task performance. The fourth perspective aimed to identify if participants' experience differed between the three UEMs.

Koutsabasis et al. defines an evaluation framework that was applied to a case study using 4 attributes [46]. The attribute realness (or relevance) refers to what degree of relevance each usability problem identified has. There are several ways to determine the relevance of a usability problem. Comparing the identified problems to a general usability problem list, have an expert review and decide the relevance of each problem or let the end-users review and decide the relevance of each problem. Validity (or accuracy) is defined as the how many times the usability problem is identified during the evaluation, compared with the total number of usability problems identified. Thoroughness (or completeness) is defined by the number of usability problems identified by the UEM in regard to the total numbers of actual problems that exist in the system evaluated by the UEM. Effectiveness has by previous research been equated to combining validity and thoroughness. Some studies take the effectiveness criteria even further, addressing the issues with the uptake of identified usability problems by the developer team. This perspective of effectiveness handles the nature of identified usability problems itself, for example that the objective usability problems are far more likely to be fixed by the developer team than the identified subjective usability problems.

Vilbergsdottir et al. [82] state that a part of analysing a usability problems validity is to prioritise what usability problems to fix and in what order. Tools to prioritise are marginal effects, budgetary constraints and frequency of problems. In a case study where two developers assessed the validity of the usability problems, they were interested in seeing the identified problems' frequency, since this information provides a quantitative picture of how many users or experts experienced a specific problem. Previous research has used a usability problems frequency combined with its severity rating to what see what impact the usability problems have [34,38,41].

Hornbæk states that assessment of formative UEMs have issues regarding validity, reliability and practical utility due to limitations in statistical testing, conclusions passed on by practitioners and the measures that are used to compare usability evaluation methods [33]. Hornbæk extends this critique, presenting seven dogmas in UEM work. The term dogma refers to a set of beliefs that is accepted without any questions from the society. The aim with the notion of these dogmases is to describe why and in what way research and evaluation of UEMs are problematic.

According to Hornbæk, the dogmas would allow a more direct approach than existing methods in regards to how to move beyond the problems. The seven dogma presented in the study are;1. Problem counting as the main approach to assessing UEMs; 2. Matching problem descriptions is straightforward; 3. Usability evaluation proceeds as prescribed and directly identifies problems; 4. The individual usability problem as the unit of analysis; 5. Look at evaluation in isolation from design; 6. Single best UEM exists & 7. Usability problems are real.

Examples of limitations that derive from the first dogma problem counting are that the relevance of the problems detected from each UEM is not assessed. Counting the problems their respective value in regard to generality, type and clarity is not considered, this leads to a problem that is considered smaller is counted equal to a problem that is considered larger. Furthermore, counting problems means that the observations made by the evaluators throughout the process give far more insights. Insights that will not be included in their documented list of usability problems. Hornbæk´s study concludes three aspects to extend the current practice of evaluating UEMs [28].

1. The problem counting should be combined with an analytical method to analyse the relevance of the identified usability problem.
2. Identify and apply methods that enable the evaluators to suggest how to implement the solutions regarding the design aspects.
3. The evaluators' insights and satisfaction with the evaluated UEMs should be considered and documented.

Since there is no formal evaluation framework that could be implemented when evaluating UEMs, the evaluation framework adopted in this study are based on the findings from the studies presented in this section.

2.4 Evaluation Framework

Identifying relationships between variables is one of the most common objective to study in the field of HCI [47]. Frøkjær et al. state that correlation analysis is important to consider since it could reveal answers and relationships that otherwise would have been missed [21]. Investigating whether the three criteria that is included this evaluation framework have any correlation is of interest since a correlation between them would indicate that all three criteria should be accounted for when evaluating an UEM.

2.4.1 Criteria Relevance

Wixon state that only counting the usability problems is not a sufficient method to use when evaluating a UEM, he argues that the problems detected needs further elaborating and assessment to understand the true value of the problem detected [85]. Furthermore, he argues that UEMs should be evaluated using the criteria success, meaning that the usability problems detected should be valued depending on how much the system or interface would improve when the problem was fixed. Van den Haak et al., stated that relevance of the usability problem was defined by how much solving a particular problem would enhance the usability of the tested product [79]. This perspective aligns with Wixon´s [85] criteria success, regarding how effective the usability performance will increase once the solution for the problem has been implemented [85]. Koutsabasis et al., concluded that evaluating how effective usability problems detected by the UEM´s here effectiveness is built upon the realness, validity, and thoroughness of the usability problem [46]. The attribute effectiveness has similar qualities and values as the attribute success stated by Wixon , relevance of the usability problem stated by Van den Haak et al., and the attribute effectiveness stated by Koutsabasis et al., [46,79,85]**.** Furthermore, relevance of usability problem as an criteria for evaluating UEM is supported by Hornbæk, who argues that evaluating UEMs by counting usability problems and not their respective value results in missing context of the problem, how relevant the problems that are detected

by the UEM actually are [33]. A criteria when evaluating UEMs should therefore assess each detected usability problem´s relevance, not only counting how many usability problem the UEM uncovered [25,33,46,85].

2.4.2 Frequency

Wixon states that a criteria that should be used when evaluating UEMs is credibility, to maintain credibility the source of the information should be verified and its trustworthiness, expertise and credentials should be validated [26,85]. Van den Haak et al., discuss the importance of UEMs' validity as well and Koutsabasis means that the validity of the usability problem can be measured by how many times a specific real problem has been identified compared to the total number of usability problems identified by the UEM [46]. Vilbergsdottir et al. [82] analyse a UEM's validity through a theory that explores how a usability problem is understood and how it is prioritised. They argue that one way to prioritise the usability problems is to measure the problems frequency. In their study, they define frequency as *The number of times a usability problem is experienced or predicted by the evaluators*. Other studies have used a usability problem's frequency and given it a severity score to calculate the problem's impact and validity [18,34,38,41]. Koutsabasis and Hvanneberg both state that validity is defined as how many of the identified usability problems are real usability problems divided by all the identified usability problems (or Validity= Hits/Hits+False Alarms) [34,46]. However, to calculate if usability problems are real, it requires either a careful cross-examination of several UEMs or extensive user testing [34,46]. Hence, a problem's frequency value affects the usability problem's validity, meaning that a usability problem's frequency value on its own is an indication of its validity. Furthermore, validity is an attribute that assessments of UEMs often miss, due to limitations regarding statistical testing and conclusions that must be made by the evaluators. Therefore, an argument can be made that calculating a usability problem's frequency can indicate the UEM's validity, a method that does not require the same number of resources as calculating the validity according to Hartson, Hvanneberg and Koutsabasis [27,33,82].

2.4.3 Criteria Timeliness

Wixon´s criteria timeliness has similar meaning to Koutsabasis et al. criteria effectiveness, which combines validity and thoroughness, they further motivate that the more relevant and of an objective nature a usability problem has, the more it is likely to be addressed by the developer team [46,85]**.** In the fifth dogma Hornbæk address that one of the reasons that many evaluations of UEMs fail is because they do not take the evaluators notes and insights from the evaluation [33]. Furthermore, he argues that the evaluators' insights ideally should be able to turn into suggestions of solutions in the design of the application being evaluated. This perspective connects to both Wixon´s opinion on usability evaluation being fundamentally flawed and lacks relevant to applied usability work and Koutsabasis et al. emphasis of the importance that the usability problems are able to be addressed by the developing team [46,85].

3 RESEARCH DESIGN

3.1 Research problem

The literature and studies on comparing usability methods is scarce, the little research that has been done on the subject are varying in quality [79]. Hasan et al. **s**tates that the existing studies that compare usability evaluation methods identified which method was more effective by comparing the number of identified problems they

produced [28]. However, the studies offered little to no explanation or analysis of the potential benefit or drawback a specific problem had. Koutsabasis et al. argues that the value of comparing usability evaluation methods is essential because it allows practitioners to have a consolidated understanding of the methods, based on multiple usability evaluation methods [46]. Furthermore, the validity and reliability from the results of the comparative studies have been difficult to confirm due to the lack of standard criteria of comparison and stable standard processes for evaluating usability evaluation methods [27]. Wixon argues that the underlying problem to evaluate usability evaluation methods is because a scientific approach to evaluating usability evaluation methods is discordant with the product developments underlying philosophy and context [85]. Due to this it is difficult to capture critical factors of the evaluation, such as the credibility and timeliness of the collected usability data.

Weichbroth´s concluded that despite the vast research regarding usability applied on mobile applications, studies regarding UEM´s within the context of mobile applications contain vague terminology [84]. The study states that there is a need to explore and refine usability evaluation systems that apply to mobile applications. Since HE is a cost effective method to uncover a large amount of usability problems, using novice evaluators such as end-users as evaluators of a system allows smaller organisations to evaluate their usability without hiring usability experts [11,29,51,71].

Furthermore, previous research about domain experts and usability expert's studies concluded that domain experts in testing environments have had difficulties identifying usability problems that are related to the interface, and usability experts have had difficulties identifying domain and task related usability problems [29]. Paz et al found that double experts (both domain and usability expert) were preferred to use when conducting a HE. Hence, it is of interest to explore how the evaluations content differ.

3.2 Research questions

- How does a HE performed by users compare with a HE performed by experts based on the evaluation criteria relevance, frequency, and timeliness?
- In each evaluation's individual result, what type of correlation exists between the criteria relevance, frequency and timeliness?

3.3 Research Paradigm

Venkatesh [73] suggests that using mixed methods as an approach will aid the researcher in acquiring a deeper understanding of complex organisational and social phenomena. Although mixed methods by some researchers have been named the third methodological paradigm, this study will combine two methodologies, several researchers have declared it is possible for multiple methodologies to coexist peacefully [81,88]. Therefore the research design for this study will assume the critical research paradigm and constructivism paradigm with the aim of developing causal explanations in the specific constructed reality of the elements studied in this thesis [5,9,17,59]. The critical research paradigm is an appropriate paradigm of choice since this study challenge existing status que in society by questioning whether heuristic evaluation only is to be conducted by experts [59]. The constructivism paradigm is appropriate to use since this study constructs its own understanding and experience of HE by conducting two evaluations which later will be reflected upon through the three evaluation criteria [3]

This mixed method study will analyse the data produced from the HE done in the case study concurrently, incorporating the combined result from both the quantitative and qualitative analysis [81]. Presenting the result from both quantitative and qualitative methods allows the analysis of the study to be presented in a unified body

that provides a holistic explanation of the phenomena [15]. Furthermore, Iorache and Pribeanu concluded that quantitative measures alone are not descriptive enough to assess the individual usability problems of a UEM [36]. They suggest an integration of both quantitative and qualitative methods, which makes a mixed method approach appropriate for this study.

3.4 Research Strategy

Employing a case study within the field of HCI is appropriate when the goal is to collect in-depth data and evaluate interfaces [47]. This research study will adopt a case study approach to evaluate the UEMs. A case study is practical to use when producing the data for applied usability [85]. This approach will ensure that the rich context surrounding the case is sufficient enough to provide the nuanced data needed to develop a broad, differential and contextualised understanding of the UEMs that are to be evaluated. Furthermore, Koutsabasis et al. argues that employing a case study when performing a comparative usability evaluation provides in depth insights about the UEMs that are evaluated [1,46].

3.4.1 Ethics

Ethics to consider when collecting user feedback on a mobile application through HE are informed consent, participants anonymity and how the collected data will be stored [20]. To ensure that the ethical aspects are respected all participants signed a consent form, the consent form includes the purpose of the research, information regarding anonymity, how the data was processed and stored (A4). All evaluations were combined into one anonymous document.

3.5 Data Collection – Heuristic Evaluation Protocol

To answer the research question two separate HEs were performed on the same application. The first evaluation was performed by five end-users of the system and the second evaluation was performed by the five usability experts. All evaluations were conducted on site. The evaluations consisted of a walkthrough were the evaluators identified usability problems, notes were taken by the moderator of each identified problem. After the walkthrough was finished the moderator went through the 11 heuristics and the evaluators connected each identified usability problem to the heuristics they though the problem belonged (A3) [53]. HE protocols usually includes five phases, planning, training, evaluation, discussion [66,80,89]. Some studies have created a formal document based on the result from the evaluation and discussion phases [55,68].

3.5.1 Planning

During the planning phase it is expected to choose what type of evaluators suits the evaluation, how many evaluators should be a part of the evaluation and what set of heuristics that are appropriate to use for this specific evaluation context [66,68,89]. Since this study is evaluating the usability of a mobile application Pimentos 11 heuristics developed for mobile applications was used [31,66,78].

To select participants for the evaluation performed by the end-users, a questionnaire was conducted (A5). The questionnaire was sent out to the applications 100 users. The questionnaire described the aim with the research and how the data would be collected. Those who were interested to participate left their contact information in the questionnaire. After two weeks there were 38 replies to the questionnaire and 24 out of those had left their contact

information. After discussion with both Vida and Skogforsk, five users that used Vida WQP with the same frequency were selected to be evaluators in the HE performed by users.

To select participants for the evaluation performed by experts five experts were selected through convenience sampling to conduct a traditional HE of Vida WQP. All experts were master students in human-computer interaction and fulfil the requirements stated by Botelli et al., Gulliksen et al., and Nielsen [10,24,63]. Nielsen argued that five evaluators is enough to uncover 75% of a systems usability problems. Hence, 5 out of the 24 harvest operators that use Vida WQP on a daily basis were selected to conduct a HE of Vida WQP as domain/novice experts [31,50,51,71,77].

3.5.2 Training

The training phase is considered important because if the evaluators lack understanding of what is meant by each heuristic and what domain the evaluation takes place there is a risk that important usability problems are missed, previous research concludes that training has an impact on the quality of the HE [40,41,32,73].

For the evaluation performed by the users, the training session consisted of a document containing the 11 heuristics with attached descriptions and applied examples, which was sent out two days before the evaluation (A2, A3). The instructions given were to read them through and ask the moderator questions if there was something that they had trouble understanding. Before starting the evaluation performed by the end-users, the moderator made sure that they understood the heuristics. For the evaluation performed by the experts, the training session consisted of distributing the same document that was given to the users 2 days prior to the evaluation. Before starting their evaluation, a short walkthrough of the app´s functionalities were explained, this was needed for the experts to grasp what type of information each functionality displayed.

3.5.3 Evaluation

During the evaluation phase each evaluator should examine the system product individually and decide whether the user interfaces (UI) are following the usability heuristics [68,78,80,89]. Depending on which study and protocol the HE chooses to follow, the evaluators could either use the application and its interface freely or they could follow predefined tasks, this study will use the latter approach [34,48,68,78].

The evaluation performed by the users took place directly after the walkthrough of the heuristics. Since the evaluators have become familiar with the heuristics during the training, instead of identifying usability problems heuristic by heuristic they identified usability problems as they visited every page in the app. When that was done revisited the heuristics and connected each problem to one or several heuristics The evaluations performed by the experts had the same structure by identifying usability problems as they did a walkthrough of the entire app.

3.5.4 Discussion

The purpose of discussing the identified usability problems as a group is to go through the list of identified problems and establish if the identified problem indeed is a usability problem [2], the discussion should address how many evaluators identified the same usability problem and if said problems have the correct description [78]. However, due to the limited time resources set for this study, it was not possible to conduction discussions with either of the two evaluation groups.

3.5.5 Formal report

The usability problems, their descriptions and the suggested solutions to each problem were included in a final formal report [62,[66,69]. Two formal documents per evaluation were produced, one document contained all identified problems (A6, A8) and the other two documents contained the unique identified problems for respective evaluation, a relevance value, a frequency value and a solution value (A7, A9).

3.6 Data analysis

3.6.1 Quantitative analysis

Quantitative analysis is appropriate to use on a large set of usability problems that are derived from a HE [49]. Lazer concludes that methods suitable to compare user studies with multiple objectives are specific significance tests such as anova tests and t-test [47]. Methods suitable to test the variables correlation or relationship to each other are logical and linear regression and contingency tables [94].

Identifying correlation between the variables means that the study is able to explore underlying relationships between the variables [47]. This is of interest since relationships between variables would indicate that some of the evaluation criteria would be appropriate to use together in future evaluations of UEMs [21]. To answer this question a chi-2 test with independent samples is used to test if two categorical variables are independent or if they are relational [42]. Independent samples of data that do not reveal any information about each other and can exist without one another are independent samples of data. Samples of data that have an influence in the other data samples are dependent samples of data [47]. Since the variable USolution value and ESolution value both depend on their respective frequency variables and are therefore not suitable to take part in a chi-2 test [7]. Appropriate analytical methods to use instead are regression analysis. Since the purpose of this analysis is model construction, meaning to answer if the independent (UFrequency and EFrequency) and dependent (USolution value and ESolution value) have a relationship [47]. Logistic regression is used for ordinal variables that contain more than two categories and for ordinal variables that only contain two categories linear regression is adequate [73].

Quantitative analysis using descriptive information and correlation tests will be conducted on the three criteria relevance, frequency and timeliness in each criteria that was generated and added in each of the evaluations formal documents (A7, A9).

3.6.2 Qualitative analysis

Computer aided content analysis, existing in the realm of computer assisted text analysis, refers to a range of techniques that derive from quantitative social science to data-mining and text-mining [12]. Using text-mining as a smart digital tool for analysing and exploring the content of each evaluation to identify common themes allows researchers to work with the content in an efficient way [72]. The term mining refers to extracting something valuable such as information and intelligence from a larger pool of less valued data. Examples of what structures and information text-mining can produce are sentiment analysis, text visualisation and topic-modelling [13]. Although text mining has a strong quantitative focus, Yu et. al argues that several researchers view text-mining as a feasible qualitative research method [87].). Even tough text mining originally was developed to handle large corpuses, several researcher claims that applying text mining on a small corpus, still returns interesting and valuable insights [83,91]. The formal protocols generated from the HEs contain semantic information which this study aims to explore, in order to see how the two evaluations differ, regarding what type of problems they tend to

identify. Therefore, the choice to use the text mining tools word frequency, word cloud and topic modelling in this specific scope is motivated.

Text-mining can derive word relationships, word frequency, themes, positive and negative enforced words in a semantic analysis [30]. Some of these produce data that a qualitative coding of information would produce [60]. The first step in text mining is to clean the data set, this was done through a function in the software Orange, where a document with relevant stopwords was added and only the columns, *Problem, Description* and *Solution* was selected for further analysis as the other column did not contain meaningful information for this analysis [52]. After the data was pre-processed, the functions bag of words and word cloud was applied. The output of these functions displays each word´s weight (frequency).

Word clouds and word frequency only provide results that are out of context, to identify word co-occurrences and themes the function topic-modelling is applied [12,54]. Topic-modelling is a text-mining tool that is useful to discover themes, frames or latent content in the documents being processed, this is based on clusters of words and their respective frequency [37,64]. Topic-modelling has in recent years assisted researchers in analysing user-generated text. Nikolento et al. argues that one obstacle for using topic-modelling and the specific algorithm LDA, is that the algorithm lacks the capabilities to identify interesting topics that match the potential to be interesting for a researcher. However, Jacobs and Tschötschel argue that the output topic-modelling produces topics and word relationships that can be interpreted and further analysed [37]. Furthermore, the abstract themes that is the LDA output can be labelled by someone possessing the knowledge of the corpus that the data has been extracted from [95].

The qualitative analysis was based on each formal protocol from the unique identified problems (A7, A9), which were uploaded in the text mining tool Orange. The methods in the qualitative analysis used the problems name, description and solutions, found in the formal protocols to generate the word frequency, word cloud and abstract themes from the topic modelling.

4 RESULT AND ANALYSIS

This section will demonstrate the quantitative results of each unique problem's value as well as the qualitative results of each evaluation's unique problems and their description and solution.

The evaluation performed by five users resulted in 97 identified problems (A6), 63 of the problems had a corresponding solution and 47 problems were unique (A7). The identified problems that had the highest frequency are displayed in image 1.

UFrequency	UProblem ID	ULocation	UProblem	UDescription
5	13	Fördelningsgrad inmätt/skördat	Missing comparable function	Difficult to compare the figures of the distribution rate from harvested and measured
5	34	Manuella kap	Red light frustration, user feedback	Finds this feature redundant, never uses. Only frustration when it is always red
5	43	Startsida/Meny	Double function	Uses the menu to enter the functions
5	44	Vrak	Missing information, data variables	Difficult to get a clear picture of wrecks in volume only

Image 1. Usability problems with the highest frequency from the evaluation performed by users.

The evaluation performed by the experts resulted in 134 identified problems, 112 of the problems had a corresponding solution (A8). Out of the 134 problems, 85 of them were unique and all the unique problems had corresponding solution/s (A9). The identified problems that had the highest frequencies are shown in image 2.

EFrequency	EProblem ID	ELocation	EProblem	EDescription
5	3	Filter	Misplaced buttons	It is confusing that the reset button is the biggest and that the save button has the same design as 7 & 30 days. On instinct you want to use the reset button as save button
5	33	Fördelningsgrad skördat, inmätt and Momdulträff, sawmill	Unclickable cards	Difficult for the user to understand how they got the next page on intuition alone. The first instinct is to click on the card and nothing happen
5	75	Startsida/Meny	Double functions	It is ok to have the same features in your interface but not the exact same thing
4	20	Fördelningsgrad inmätt/skördat	Missing comparable function	The users have to remember the results from fördelningsgrad inmätt, in order to compare these with fördelningsgrad inmätt.

Image 2. Usability problems with the highest frequency from the evaluation performed by experts.

Since user evaluation generated 47 unique problems and the expert evaluation generated 85 unique problems it is motivated to use their descriptive statistics to see how they compare to each other and explore the correlation between the variables using statistical tests such as chi-2 test and linear and logical regression [47].

4.1 Quantitative results

A Sharpio-Wilk p test was conducted on all six variables to check if the data was normally distributed, all variables had a p-value of <0.001 indicating that the data is normally distributed [47].

4.1.1 Frequency

Frequency, which is an indication of a problem's validity, measured how often the same problem occurs by counting how many times a problem occurs out of the total number of problems identified [18,34,38,41,82]. In the evaluation performed by the users, the mean and standard deviation value of the variable frequency were 1.94 and 1.28. In the evaluation performed by the experts, the mean and standard deviation of the value frequency were 1.59 and 1.03 (image 3). This means that the user evaluation has a higher frequency value, which indicates that they have more validated usability problems than the expert evaluation.

This result indicates that the number of identified problems did not matter for this criteria, seeing that both evaluations had a similar mean value and that the user evaluation had a slightly higher value than the expert evaluation had. The users slightly higher mean-value could be the result of their homogeneity as a group. This could have affected the frequency value. If the end-users that participated in the evaluation would have had varying interest and usage of the app, this could have affected the frequency and produced a lower value.

4.1.2 Relevance

Relevance is measured by a system developer who is considered an expert within the system that is being evaluated. The system developer gives each individual usability problem a value between one and five, depending on how feasible the problem is to practically implement [25,27,33,46,56,85]. In the evaluation performed by the users, the mean, median and standard deviation value of the variable Relevance were 3.23, 3 and 1.29 In the evaluation performed by the experts, the mean, median and standard deviation of the value Relevance were 3.18, 3 and 0.966 (image 3). This means that the user evaluation has a higher relevance rate than the expert evaluation.

This result indicates, as stated above, that the number of problems did not appear to affect the results since both evaluations produced similar results, even though the user evaluation from these results is considered to have a highest value, and therefore problems with higher feasibility these results were based on a single system developers' expert opinion. Furthermore, given the developer works closely with the end-user, one can assume they have more similar opinions and experience with the system than the usability experts. Changing or adding other experts from different fields to rate the feasibility of problems would most likely change the outcome of this result.

4.1.3 Timeliness

Timeliness is measured by counting the number of solutions that evaluators give to each unique problem and dividing them by their frequency attribute, which produces the variable solution value [23,33,34,46,84,85]. In the evaluation performed by the users, the mean, median and standard deviation value of the variable *Solution value* were 0.723, 1 and 0.396. In the evaluation performed by the experts, the mean, median and standard deviation of the value *Solution value* were 0.8, 1 and 0.402 (image 3). The results show that the expert evaluation has a higher solution rate than the user evaluation.

This result indicates, as identified in the two previous results, that the number of problems did not appear to affect the results, since both evaluations performed similarly. This result is merely based on the number of solutions, which in one way could be limited as it only counts the number of problems. To get an in depth and

nuanced understanding of this criteria in future work, the solutions frequency (and if possible its validity) and relevance should be rated as well.

Descriptives

	EFrequency	UFrequency	Esolution value	USolution value	ERelevance (1-5)	URelevance (1-5)
N	85	47	85	47	85	47
Missing	0	38	0	38	0	38
Mean	1.59	1.94	0.800	0.723	3.18	3.23
Median	1	1	1	1.00	3	3
Standard deviation	1.03	1.28	0.402	0.396	0.966	1.29
Minimum	1	1	0	0.00	1	1
Maximum	5	5	1	1.00	5	5
Shapiro-Wilk W	0.632	0.739	0.490	0.685	0.902	0.897
Shapiro-Wilk p	< .001	< .001	< .001	< .001	< .001	< .001

Image 3. Results of each measured variable for both evaluations performed by users and experts.

4.1.4 Correlation tests

The significance level of all statistical tests has been set to 0.05. If the p-value of the chi-2, linear and logical regression tests are below 0.05 it means that the variables are not independent and that there is a relationship between the variables. Thus, the null hypothesis (H0) for the correlation tests have been formulated as ´There is no significant relationship between the two independent variables`. The result of the chi-2 test on the variables UFrequency/URelevance had a p-value of 0.745 (A10.1). The chi-2 test for variables EFrequency/ERelevance had a p-value of 0.099 (A10.2) and the H0 has been accepted in this case. EFrequency/ERelevance. The result of the chi-2 tests on the variables USolution value/URelevance had a p-value of 0.374 (A11.1) and the H0 has been accepted. The chi-2 test for variables EFrequency/ERelevance had a p-value of 0.665 (A11.2) and the H0 has been accepted in this case as well. From these results we can conclude that there is no significant relationship between the variables UFrequency/URelevance, USolution value/URelevance, EFrequency/ERelevance and ESolution value/ERelevance.

The result of the logistical regression for variables UFrequency/USolution value had a p-value of 0.180 (A10.3) and the linear regression for variables EFrequency/ERelevance had a p-value of 0.188 (A11.3), and the H0 has been accepted in both cases. The statistical correlation analysis concludes that there are no relationships between the criteria relevance, frequency and timeliness; each criteria is independent and have no influence on each other. As stated before, exploring the variables' correlation was of interest to see if they statistically could be argued to use all three criteria as a formal evaluation framework [22]. Even if this result does not show any correlation between the variables, revisiting this question with a larger dataset is still of interest, as one could argue that the 47 unique problems generated by the users is on the verge of what is suitable to apply quantitative analysis methods such as correlation tests on.

4.2 Qualitative results

The criteria timeliness also explores the meaning in the content, taking on a qualitative approach [33]. The qualitative results in this study were generated from the two evaluations' uniquely identified problems and then uploaded in the text-mining tool Orange (A7, A9). Between the two evaluations, there is a difference in what type of words had the highest frequency. Image 4 displays the word frequency from the two formal documents of the evaluations unique identified problems. The greater weight of a word, the bigger it is illustrated in the word clouds (A12, A13).

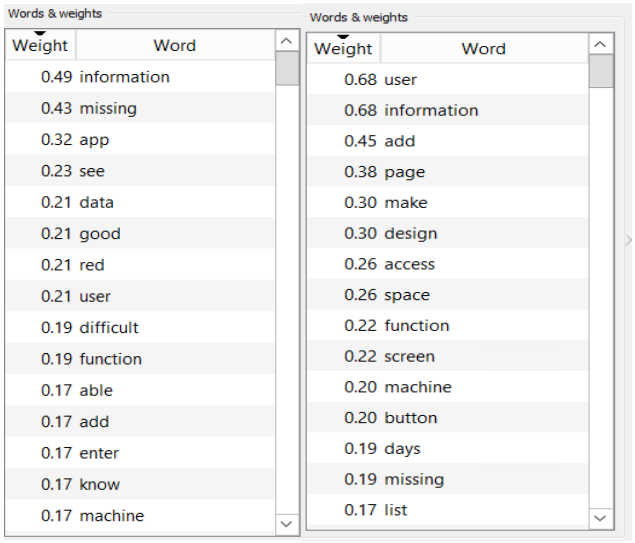


Image 4. Word Frequency-Users Word-frequency- Experts

4.2.1 User Evaluation

The ten abstract themes provided in the topic analysis from the user evaluation (A14) have received labels by the researcher that conducted, cleaned, and organised the data and therefore possess deep knowledge about the data set. The first abstract theme contained words that the user evaluators expressed regarding the navigation in the app and their feeling towards it. The second abstract theme contained words that evaluators expressed about the app´s consistency and standards. The third theme contained words that the evaluators expressed regarding the apps output and their received feedback. The fourth abstract theme contained words that the evaluators expressed in regard to how the information in each function was displayed. The fifth abstract theme contained words that the evaluators expressed in regard to what functions they thought were useful. In the sixth theme, the evaluators words were about the visual presentation of the app´s interface. The seventh abstract theme contained words in regards to the ease of use by intuition. The eighth abstract theme contained words regarding the confusions for some functionalities and the app´s lack of help and error handling. The ninth abstract theme contained the evaluators expression regarding the lack of error prevention and inconsistency in user language. The tenth abstract theme contained words that the evaluators expressed opinions considering what data variables they would want to add and variables that they would like to add descriptive data to in the app. All ten themes received labels from the

moderator that described their content. This result illustrates what type of problems that the users identified during the evaluation, most of their themes regards how information is displayed in the app, which can be seen in theme two, three, four, five, nine and ten (Table 1).

Table 1. Abstract themes from user evaluation labelled

No	Label	Abstract word collection
1.	In app navigation	Add, menu, lengths, remove, places, frustrating, two, double, press, explain
2.	Inconsistent standards	Via, dark, disappears, mode, ext, home, inconsistent standards, phone, different, measurement
3.	Feedback and observation	Know, length, user, entered, able, wrecks, feedback, machines, wrong, addition
4.	Access to information	Versions, overload, keep, correct, important, outdated, necessary, old, possibility, spruce
5.	Adequacy of core information	Bucking, frustration, parameters, light, wrecks, anything, frustrated, influence, addition, manual
6.	Visual presentation	Functions, make, purpose, difficult, lights, could, function, directly, sawmills, starts, cuts
7.	Memory load and intuitive design	Understand, purpose, difficult, lights, could, function, fully, descriptions, color, blind
8.	Adequacy of functionality, help and error handling	Colour, page, start, something, blind, choose, delete, says, confusing
9.	Error and user language prevention	App, red, uses, settings, possible, error, handling, press, explain, double
10.	Data variables and descriptive information	Information, good, missing, variables, volume, interesting, timber, data, see, enter

4.2.2 Expert Evaluation

The ten abstract themes provided in the topic analysis from the expert evaluation (A15) have received labels by the researcher, same as the abstract theme from the user evaluation. The first abstract theme contained words that the expert evaluators expressed regarding the ease of use by intuition. The second abstract theme contained words in regard to how and where the information is placed in the app. The third abstract theme contained words that regarded the navigation in the app and their feeling towards it. The fourth abstract theme contained words that the evaluators expressed in regards to how the information in each function was displayed. The fifth abstract theme expressed the evaluators opinions regarding the app´s design. Similar to the fourth theme, the sixth abstract theme contained words that the evaluators expressed were about how the information was displayed in the app. The seventh abstract theme contained words that regarded how the app laced consistency in user language. The eighth abstract theme contained words that regarded the users feelings and reactions towards the app. The ninth abstract theme contained words that the evaluators expressed regarding the misuse of screen space, some attributes could be designed to be bigger and some places were overcrowded and should be cleaned. The tenth abstract theme contained words that evaluators expressed about the app´s consistency and standards. All ten themes received labels from the moderator that described their content. This result illustrates what type of problems that the experts identified during the evaluation, most of their themes were connected to the applications interface, this could be seen in themes one, two, five and nine (Table 2).

Table 2. Abstract themes from expert evaluation labelled

No	Label	Abstract word collection
1.	Memory load and intuitive design	Total, regarding, clarify, belongs, calculation, scroll, move, logical, consist, price
2.	Visual presentation and element of information	User, page, make, top, access, give, see, description, function, days
3.	In app navigation	Either, menu, functions, two, inställningar, logga, meny, mnu, icons, ut
4.	Access to information	Information, add, space, remove, graph, difficult, easily, result, another, descriptive
5.	Re-designing the interface	Text, screen, fit, instead, easily, misuse, choose, table, line, designed
6.	Access to information	Displayed, two, screens, percentage, read, misslyckade, overview, lyckade, gran, separate
7.	User language	Everything, buttons, user language, correct, correctly, Swedish, English, spelled, objects, sure
8.	Adequacy of information and functionality	Time, feels, periods, clear, different, view, money, interesting, informations, look
9.	Inconsistent use of screen space	Unclear, could, design, use, lot, know, numbers, something, percentage total
10.	Inconsistent standards	Headlines, headline, header, similar, designed, line, guide, app, visual, calculation

5 DISCUSSION

The aim of this study was to seek answers as to whether a HE performed by users, instead of experts for who the method was developed for, could have comparable results [16,32,50,62,63,66]. This study was conducted within the critical research and constructivism paradigm. The critical research paradigm was appropriate to use since this study challenged the social and normative phenomena that heuristic evaluation should only use experts as evaluators [5,9,17,59]. Employing the constructivism paradigm together with the critical research paradigm allowed this study to successfully understand both the evaluation criteria and the UEM heuristic evaluation [81,88].

Deconstructing the related work regarding evaluating UEMs and constructing three evaluation criteria resulted in an in-depth understanding of the criteria and how they were to be implemented. A similar procedure was performed on the UEM, heuristic evaluation, which was deconstructed by taking apart the HE in four different pieces: planning, training, evaluation, and formal protocol. From these four pieces, two HEs were constructed. Combining the two paradigms resulted in an in-depth understanding of the social constructions heuristic evaluation and the three evaluation criteria: relevance, frequency and timeliness. The process of deconstructing both the evaluations and their criteria, and then constructing them in the context of where they belong allowed the attributes of this thesis to continuously be reflected upon from start to finish, which helped the thesis to fulfil its aim of answering the research questions of how the two evaluations were compared.

5.1 How does a HE performed by users compare with a HE performed by experts based on the evaluation criteria relevance, frequency, and timeliness?

In previous research regarding what criteria the evaluation of a UEM should be based on, relevance of each problem, instead of the number of problems detected could be found in the majority of the articles identified in the literary study [27,33,56,79,85]. As stated by previous researchers, there lies a higher value in usability problems that have high feasibility. From the result of the descriptive statistics, the evaluation performed by users had a higher mean value for the relevance criteria. However, these results are dependent on the one system developer that worked closely with the end-user since that person rated the usability problems relevance, which ultimately

could have biased the result because they have a relationship with each other, which is something that the experts did not have.

Frequency, which is an indication of the validity was emphasised to be an important criterion when evaluating an UEM [26,33,46,79,85]**.** From the result of this study, the evaluation performed by the users had a higher mean value of the criteria frequency. Although the users by definition and scope of this study were classified as novice evaluators, one could argue that the reason for this was that they could also be classified as domain experts, since they are experts of the content and purpose of the application [66]. This indicates that users' extensive knowledge, experience and enthusiasm within the domain of the WQP application contributed to the high frequency value [29].

The criteria timeliness evaluates an UEM on different criteria for identified problems, such as solution rate and type of problem [23,33,34,46,84,85]. Hornbæk and other researchers emphasises the importance of evaluating the solutions provided by the evaluators for each identified usability problem.

The quantitative results of timeliness were that the expert evaluation had a higher solution value compared to the user evaluation, which means that the experts had provided more solutions to the number of usability problems they identified. However, since this attribute is only based on the number of solutions they produced, to explore what type of problems and solutions that were identified from a timeliness perspective, a qualitative approach utilising text-mining as a method was appropriate [30,87]. The word frequency illustrated in image 4 reveals that there is a clear difference in what the words imply for each evaluation. The user's most frequent word contains subjective words such as “good”, “able” and “difficult”, the other words could be connected to task related problems, such as “information”, “missing”, “data” and “machine”. The experts' most frequent words contain words that could be connected to the interface of the system with words such as “page”, “design”, “access”, “space”, “button” and “screen”.

Furthermore, the ten themes from the user evaluation provided by the topic-modelling primarily resulted in *Access to information*, *feedback*, *how to navigate in the app*, and different approaches that regarded *error handling* and *help* in how the application worked (Table 1). The ten themes generated from the expert evaluation partly contained the same themes as the user evaluation (Table 2). The themes the evaluations shared were *Memory load and intuitive design, Inconsistent standards* and *Access to information*. The difference in the content followed the same pattern as shown in their word frequencies.

The themes generated by the user were mostly focused on task related difficulties, for example they identified several data variables that were missing which are important for receiving the correct information regarding the measurements of each functionality, confirming Hassan et al. statement of what type of usability problems the two expert groups tend to identify [29]**.** Whereas the expert evaluation generated themes with specific connection to the interface, for example they identified buttons with confusing design and placement and several pages where the screen space was not properly used further confirming previous statements regarding what type of usability each expert evaluating group identified [29,67,74,86,89].

Even though the evaluations results and content differ, their quantitative results still were pretty similar and, in that sense, they actually compare on an even level. The differences in what type of problems they identified indicate that the type of evaluators should be chosen based on the type of problem that the evaluations’ aims to identify and what resources are set for the evaluation. These results are interesting finds because they indicate that depending on what the aim of the HE is, different types of evaluators should be considered. For example, performing a HE where the budget is limited and the goal is to evaluate the systems functionality rather than its interface, recruiting the systems end-users to be evaluators would be more beneficial than recruiting usability experts.

Furthermore, these results also indicate that if the aim is to uncover as many usability problems as possible it could be more important to include a variety of evaluators rather than the sheer number of evaluators.

5.2 In each evaluation's individual result, is there any correlation between the criteria relevance, frequency and timeliness?

This study concluded that there was no relationship between the evaluating variables, since none of the variable had a significance lever below 0.05. If there would have existed a correlation between the evaluation criteria, it would have been interesting to greater explore their behaviours, since currently these results mean that neither of the three variables act according to each other [47]. Therefore, it is possible to not use all three criteria when evaluating UEMs.

6 CONCLUSION

The approach mixed methods resulted in an in depth and high-quality analysis outcome. This study has provided insights regarding how an UEM can be evaluated through three criterias that are based on previous research within the field of HCI. Furthermore, valuable insights whether users as novice evaluators would identify the applicable and comparable usability problems as expert evaluators would do has been concluded.

From the result of the two evaluations that were conducted, the organisation Vida AB received two formal documents from both HEs. These evaluations provided Vida AB with useful insights regarding their application WQP´s usability. From the formal documents three variables were calculated to measure relevance, frequency and timeliness, the mixed methods approach aided the appreciation of both the quantitative and qualitative variables. The quantitative analysis generated each evaluation a concrete value for each variable and the qualitative analysis offered valuable insights in what type of problem each evaluation generated.

Since this study employed homogeneous end-users as evaluators the result could be generalised compared to previous studies that have used specific profiles as novice evaluators. Based on the quantitative result the evaluation performed by users had a higher frequency value and higher relevance compared to the expert evaluation. Furthermore, this study can acknowledge that both domain and usability evaluators identify different types of usability problems, and it therefore is beneficial to either have double experts or engage a mix of domain experts and usability experts in the HE. This study suggests that conducting a HE with end-users as evaluators will return useful and valuable knowledge about the current state the usability of a system is in.

This study can conclude that having users as the evaluators in a HE can produce relevant and valid results, and with this unlocking an inclusive, simple and cost-effective way to evaluate the usability of a system.

7 LIMITATIONS AND FUTURE WORK

There is no widely accepted evaluation framework, thus the framework for this study is based on a related field within this topic [27,28,79,84]. Although thorough research was conducted to build a framework, some aspects when evaluating UEMs could have been missed. Due to the inaccessibility to communicate over time with the end-users, no final group discussion was held, therefore there was no discussion conducted with the experts as evaluators, which might have affected the result of each evaluation because the aim of a discussion is to see if all the evaluators agree with the usability problems, their descriptions and their solutions that are listed [2,78]. Due to this, the discussion could have affected what problems that the formal documents would have contained. The criteria relevance was based on the one system developer expert from Vida AB´s and knowledge. If more experts

and experts with experience in designing user interfaces the result of the criteria relevance might have a different outcome than this study had [33,46,76,79,86]. Validity was said by several previous researchers to be an important criteria when evaluating and comparing UEMs [33,34,46,85]. Due to the lack of time and access to end users, this study could not conduct the number of UEMs needed to identify what usability problems were considered to be real problems. Although a usability problem's frequency is an indication of a problem's validity, it cannot be considered equal to how validity has been defined by previous researchers, since this study has not verified which identified usability problems are real and which are false. This means that some of the usability problems identified in this study might not be real problems, which would affect the UEM's validity, and might have a different outcome than what the frequency criteria showed [34]. Future work should therefore implement a proper validity scheme, for example by cross-examining several UEMs identified usability problems to determine what usability problems are real and what problems are false hits [27,34,46].

As suggested by several studies, a case study is the most effective way to evaluate and compare UEMs [1,46,47,85]. Within the field of HCI results produced from case studies are considered to be generalisable when it could be applicable to a broader target group [14]. The results of the study did not produce generalisability as a whole. However, the study did use a homogenous group of end-users as evaluators for the novice HE. Which for future work can be adopted when choosing types of evaluators for a HE since the result for this group produced generalisable results [29,40,51,71,77]. Despite this, the study in its entirety cannot be regarded as generalisable, the results encourage further research and actual implementation in regards of considering end-users a form of domain experts [29,66].

REFERENCES

[1] Trudie Aberdeen. 2013. Yin, R. K. (2009). Case study research: Design and methods (4th Ed.). Thousand Oaks, CA: Sage. *The Canadian Journal of Action Research* 14, 1 (2013), 69–71. DOI:https://doi.org/10.33524/cjar.v14i1.73

[2] Roobaea Alroobaea and P.J. Mayhew. 2014. *How Many Participants are Really Enough for Usability Studies?* DOI:https://doi.org/10.1109/SAI.2014.6918171

[3] Fenio Annansingh and Kerry Howell. 2016. Using Phenomenological Constructivism (PC) to Discuss a Mixed Method Approach in Information Systems Research. *Electronic Journal of Business Research Methods* 14, 1 (September 2016), pp39-49-pp39-49.

[4] Muhammad Faisal Aziz, Harlili, and Dicky Prima Satya. 2020. Designing Human-Computer Interaction for E-Learning using ISO 9241-210:2010 and Google Design Sprint. In *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, 1–6. DOI:https://doi.org/10.1109/ICAICTA49861.2020.9429074

[5] Jose Molina Azorin and Roslyn Cameron. 2010. The Application of Mixed Methods in Organisational Research: A Literature Review. Electron. J. Bus. Res. Methods 8, 2 (December 2010), pp95-105.

[6] J. M. Christian Bastien. 2010. Usability testing: a review of some methodological and technical aspects of the method. *International Journal of Medical Informatics* 79, 4 (April 2010), e18–e23. DOI:https://doi.org/10.1016/j.ijmedinf.2008.12.004

[7] Mark P. Becker and Clifford C. Clogg. 1989. Analysis of Sets of Two-Way Contingency Tables Using Association Models. *Journal of the American Statistical Association* 84, 405 (March 1989), 142–151. DOI:https://doi.org/10.1080/01621459.1989.10478749

[8] Enrico Bertini, Silvia Gabrielli, Stephen Kimani, Tiziana Catarci, and Giuseppe Santucci. 2006. Appropriating and assessing heuristics for mobile computing. In *Proceedings of the working conference on Advanced visual interfaces* (AVI '06), Association for Computing Machinery, New York, NY, USA, 119–126. DOI:https://doi.org/10.1145/1133265.1133291

[9] Frank Bogna, Aldo Raineri, and Geoff Dell. 2020. Critical realism and constructivism: merging research paradigms for a deeper qualitative study. *Qualitative Research in Organizations and Management: An International Journal* 15, 4 (January 2020), 461–484. DOI:https://doi.org/10.1108/QROM-06-2019-1778

[10] Federico Botella, Eloy Alarcon, and Antonio Peñalver. 2014. How to classify to experts in usability evaluation. In Proceedings of the XV International Conference on Human Computer Interaction, Association for Computing Machinery, New York, NY, USA, 1–4. DOI:https://doi.org/10.1145/2662253.2662278

[11] Mike Brayshaw, Neil Gordon, Julius Nganji, Lipeng Wen, and Adele Butterfield. 2014. Investigating Heuristic Evaluation as a Methodology for Evaluating Pedagogical Software: An Analysis Employing Three Case Studies. In *Learning and Collaboration Technologies. Designing and Developing Novel Learning Experiences*, Springer International Publishing, Cham, 25–35. DOI:https://doi.org/10.1007/978-3-319-07482-5_3

[12] Alan Brier and Bruno Hopp. 2011. Computer assisted text analysis in the social sciences. *Qual Quant* 45, 1 (January 2011), 103–128. DOI:https://doi.org/10.1007/s11135-010-9350-8

[13] Hsinchun Chen, Roger H. L. Chiang, and Veda C. Storey. 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly* 36, 4 (2012), 1165–1188. DOI:https://doi.org/10.2307/41703503

[14] Parmit K. Chilana, Amy J. Ko, and Jacob Wobbrock. 2015. From User-Centered to Adoption-Centered Design: A Case Study of an HCI Research Innovation Becoming a Product. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), Association for Computing Machinery, New York, NY, USA, 1749–1758. DOI:https://doi.org/10.1145/2702123.2702412

[15] Julie A Corrigan and Anthony J Onwuegbuzie. 2020. Toward a Meta-Framework for Conducting Mixed Methods Representation Analyses to Optimize Meta-Inferences. In The Qualitative Report 25, 3 (March 2020), COV15-COV15.

[16] Ruyther Costa and E.D. Canedo. 2019. A Set of Usability Heuristics for Mobile Applications. In HumanComputer Interaction. Perspectives on Design (Lecture Notes in Computer Science), Springer International Publishing, Cham, 180–193. DOI:https://doi.org/10.1007/978-3-030-22646-6_13

[17] Emanuel Felipe Duarte and M. Cecília C. Baranauskas. 2016. Revisiting the Three HCI Waves: A Preliminary Discussion on Philosophy of Science and Research Paradigms. In *Proceedings of the 15th Brazilian*

Symposium on Human Factors in Computing Systems (IHC '16), Association for Computing Machinery, New York, NY, USA, 1–4. DOI:https://doi.org/10.1145/3033701.3033740

[18] Henry Duh, Gerald Tan, and Vivian Chen. 2006. *Usability evaluation for mobile device: A comparison of laboratory and field tests*. DOI:https://doi.org/10.1145/1152215.1152254

[19] Mingming Fan, Jinglan Lin, Christina Chung, and Khai N. Truong. 2019. Concurrent Think-Aloud Verbalizations and Usability Problems. *ACM Trans. Comput.-Hum. Interact.* 26, 5 (July 2019), 28:1-28:35. DOI:https://doi.org/10.1145/3325281

[20] Farnaz Fotrousi, Norbert Seyff, and Jürgen Börstler. 2017. Ethical Considerations in Research on User Feedback. In *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)*, 194–198. DOI:https://doi.org/10.1109/REW.2017.68

[21] Erik Frøkjær, Morten Hertzum, and Kasper Hornbæk. 2000. Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated. In *in: Proceedings of the CHI 2000 Conference on Human Factors in Computing Systems, ACM*, Press, 345–352.

[22] Erik Frøkjær and Kasper Hornbæk. 2004. Input from usability evaluation in the form of problems and redesigns: results from interviews with developers. (January 2004).

[23] Dominic Furniss, Ann Blandford, and Paul Curzon. 2007. Usability evaluation methods in practice: understanding the context in which they are embedded. In *Proceedings of the 14th European conference on Cognitive ergonomics: invent! explore!* (ECCE '07), Association for Computing Machinery, New York, NY, USA, 253–256. DOI:https://doi.org/10.1145/1362550.1362602

[24] Jan Gulliksen, Inger Boivie, and Bengt Göransson. 2006. Usability professionals—current practices and future development. *Interacting with Computers* 18, 4 (July 2006), 568–600. DOI:https://doi.org/10.1016/j.intcom.2005.10.005

[25] Maaike J. van den Haak, Menno D. T. de Jong, and Peter Jan Schellens. 2006. Constructive Interaction: An Analysis of Verbal Interaction in a Usability Setting. *IEEE Transactions on Professional Communication* 49, 4 (December 2006), 311–324. DOI:https://doi.org/10.1109/TPC.2006.885865

[26] Giannis Haralabopoulos, Ioannis Anagnostopoulos, and Sherali Zeadally. 2016. The Challenge of Improving Credibility of User-Generated Content in Online Social Networks. *J. Data and Information Quality* 7, 3 (August 2016), 13:1-13:4. DOI:https://doi.org/10.1145/2899003

[27] H. Hartson, Terence Andre, and Robert Williges. 2003. Criteria For Evaluating Usability Evaluation Methods. *Int. J. Hum. Comput. Interaction* 15, (February 2003), 145–181. DOI:https://doi.org/10.1207/S15327590IJHC1501_13

[28] Layla Hasan, Anne Morris, and Steve Probets. 2012. A comparison of usability evaluation methods for evaluating e-commerce websites. *Behaviour & Information Technology* 31, 7 (July 2012), 707–737. DOI:https://doi.org/10.1080/0144929X.2011.596996

[29] Muhammad Mustafa Hassan, Markku Tukiainen, and Adnan N. Qureshi. 2019. Participatory Heuristic Evaluations of Jeliot Mobile : End-users evaluating usability of their mlearning application. In *2019 4th Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)*, 1–6. DOI:https://doi.org/10.1109/TIMES-iCON47539.2019.9024452

[30] Emily Haynes, Ruth Garside, Judith Green, Michael P. Kelly, James Thomas, and Cornelia Guell. 2019. Semiautomated text analytics for qualitative data synthesis. *Research Synthesis Methods* 10, 3 (2019), 452–464. DOI:https://doi.org/10.1002/jrsm.1361

[31] Setia Hermawati and Glyn Lawson. 2016. Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus? *Applied Ergonomics* 56, (September 2016), 34–51. DOI:https://doi.org/10.1016/j.apergo.2015.11.016

[32] Morten Hertzum and Niels Ebbe Jacobsen. 2001. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human–Computer Interaction* 13, 4 (December 2001), 421–443. DOI:https://doi.org/10.1207/S15327590IJHC1304_05

[33] Kasper Hornbæk. 2010. Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology* 29, 1 (January 2010), 97–111. DOI:https://doi.org/10.1080/01449290801939400

[34] Ebba Thora Hvannberg, Effie Lai-Chong Law, and Marta Kristín Lérusdóttir. 2007. Heuristic evaluation: Comparing ways of finding and reporting usability problems. *Interacting with Computers* 19, 2 (March 2007), 225–240. DOI:https://doi.org/10.1016/j.intcom.2006.10.001

[35] Rodolfo Inostroza, Cristian Rusu, Silvana Roncagliolo, and Virginica Rusu. 2013. Usability heuristics for touchscreen-based mobile devices: update. In *Proceedings of the 2013 Chilean Conference on Human - Computer Interaction* (ChileCHI '13), Association for Computing Machinery, New York, NY, USA, 24–29. DOI:https://doi.org/10.1145/2535597.2535602

[36] Dragoş Daniel Iordache and Costin Pribeanu. 2009. A Comparison of Quantitative and Qualitative Data from a Formative Usability Evaluation of an Augmented Reality Learning Scenario. In Informatica Economica Journal 13 (January 2009).

[37] Thomas Jacobs and Robin Tschötschel. 2019. Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *International Journal of Social Research Methodology* 22, 5 (September 2019), 469–485. DOI:https://doi.org/10.1080/13645579.2019.1576317

[38] Fatemeh rangraz jeddi, Ehsan Nabovati, Reyhane Bigham, and Seyede Razieh Farrahi. 2020. Usability evaluation of a comprehensive national health information system: A heuristic evaluation. *Informatics in Medicine Unlocked* 19, (April 2020), 100332. DOI:https://doi.org/10.1016/j.imu.2020.100332

[39] Cristhy Jimenez, Cristian Rusu, Virginica Rusu, Silvana Roncagliolo, and Rodolfo Inostroza. 2012. Formal specification of usability heuristics: how convenient it is? In *Proceedings of the 2nd international workshop on Evidential assessment of software technologies* (EAST '12), Association for Computing Machinery, New York, NY, USA, 55–60. DOI:https://doi.org/10.1145/2372233.2372249

[40] Claudine B. Kabeza, Lorenz Harst, Peter E. H. Schwarz, and Patrick Timpel. 2020. A qualitative study of users' experiences after 3 months: the first Rwandan diabetes self-management Smartphone application "Kir'App." *Therapeutic Advances in Endocrinology and Metabolism* 11, (April 2020). DOI:https://doi.org/10.1177/2042018820914510

[41] Claire-Marie Karat, Robert Campbell, and Tarra Fiegel. 1992. Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '92*, ACM Press, Monterey, California, United States, 397–404. DOI:https://doi.org/10.1145/142750.142873

[42] Maria Kateri. 2014. Analysis of Two-way Tables. In *Contingency Table Analysis: Methods and Implementation Using R*, Maria Kateri (ed.). Springer, New York, NY, 17–61. DOI:https://doi.org/10.1007/978-0-8176-4811-4_2

[43] Kamran Khowaja and Dena Al-Thani. 2020. New Checklist for the Heuristic Evaluation of mHealth Apps (HE4EH): Development and Usability Study. *JMIR mHealth and uHealth* 8, 10 (October 2020), e20353. DOI:https://doi.org/10.2196/20353

[44] Shazeeye Kirmani. 2008. Heuristic evaluation quality score (HEQS): defining heuristic expertise. *J. Usability Studies* 4, 1 (November 2008), 49–59.

[45] Shazeeye Kirmani and Shanmugam Rajasekaran. 2007. Heuristic evaluation quality score (HEQS): a measure of heuristic evaluation skills. *J. Usability Studies* 2, 2 (February 2007), 61–75.

[46] Panayiotis Koutsabasis, Thomas Spyrou, and John Darzentas. 2007. *Evaluating Usability Evaluation Methods: Criteria, Method and a Case Study*. DOI:https://doi.org/10.1007/978-3-540-73105-4_63

[47] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research Methods in Human-Computer Interaction*. Elsevier Science & Technology, San Francisco, UNITED STATES. Retrieved May 4, 2022 from http://ebookcentral.proquest.com/lib/uu/detail.action?docID=4851896

[48] Adrian Lecaros, Freddy Paz, and Arturo Moquillaza. 2021. Challenges and Opportunities on the Application of Heuristic Evaluations: A Systematic Literature Review. In *Design, User Experience, and Usability: UX Research and Design*, Springer International Publishing, Cham, 242–261. DOI:https://doi.org/10.1007/978-3-030-78221-4_17

[49] Xiaosong Li, Ye Liu, Zizhou Fan, and Will Li. 2018. A Quantitative Approach in Heuristic Evaluation of E-commerce Websites. (January 2018). DOI:https://doi.org/10.5121/ijaia.2018.9101

[50] André de Lima Salgado and André Pimenta Freire. 2014. Heuristic Evaluation of Mobile Usability: A Mapping Study. In *Human-Computer Interaction. Applications and Services*, Springer International Publishing, Cham, 178–188. DOI:https://doi.org/10.1007/978-3-319-07227-2_18

[51] André de Lima Salgado and Renata Pontin de Mattos Fortes. 2016. Heuristic Evaluation for Novice Evaluators. In *Design, User Experience, and Usability: Design Thinking and Methods*, Springer International Publishing, Cham, 387–398. DOI:https://doi.org/10.1007/978-3-319-40409-7_37

[52] Bioinformatics Laboratory Ljubljana University of. Word Cloud. Retrieved May 6, 2022 from https://orangedatamining.com/widget-catalog/text-mining/wordcloud/
[53] Olibário Machado Neto and Maria da Graça Pimentel. 2013. Heuristics for the assessment of interfaces of mobile devices. In *Proceedings of the 19th Brazilian symposium on Multimedia and the web* (WebMedia '13), Association for Computing Machinery, New York, NY, USA, 93–96. DOI:https://doi.org/10.1145/2526188.2526237
[54] Carmel McNaught and Paul Lam. 2010. Using Wordle as a Supplementary Research Tool. *Qual Rep* 15, (May 2010). DOI:https://doi.org/10.46743/2160-3715/2010.1167
[55] Na Mi, Lora Cavuoto, Kenneth Benson, Tonya Smith-Jackson, and Maury Nussbaum. 2013. A heuristic checklist for an accessible smartphone interface design. *Universal Access in the Information Society* 13, (November 2013). DOI:https://doi.org/10.1007/s10209-013-0321-4
[56] Rolf Molich, Meghan R Ede, Klaus Kaasgaard, and Barbara Karyukin. 2004. Comparative usability evaluation. *Behaviour & Information Technology* 23, 1 (January 2004), 65–74. DOI:https://doi.org/10.1080/0144929032000173951
[57] Tina M. Morrison, Prasanna Hariharan, Chloe M. Funkhouser, Payman Afshari, Mark Goodin, and Marc Horner. 2019. Assessing Computational Model Credibility Using a Risk-Based Framework: Application to Hemolysis in Centrifugal Blood Pumps. *ASAIO Journal* 65, 4 (May 2019), 349–360. DOI:https://doi.org/10.1097/MAT.0000000000000996
[58] Karima Moumane, Ali Idri, and Alain Abran. 2016. Usability evaluation of mobile applications using ISO 9241 and ISO 25062 standards. *SpringerPlus* 5, 1 (April 2016), 548. DOI:https://doi.org/10.1186/s40064-016-2171-z
[59] Ferdinand C. Mukumbang. 2021. Retroductive Theorizing: A Contribution of Critical Realism to Mixed Methods Research. *Journal of Mixed Methods Research* (December 2021), 15586898211049848. DOI:https://doi.org/10.1177/15586898211049847
[60] Laura K. Nelson, Derek Burk, Marcel Knudsen, and Leslie McCall. 2021. The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods. *Sociological Methods & Research* 50, 1 (February 2021), 202–237. DOI:https://doi.org/10.1177/0049124118769114
[61] Jakob Nielsen. 1994. *Usability Engineering*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
[62] Jakob Nielsen and Thomas K. Landauer. 1993. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (CHI '93), Association for Computing Machinery, New York, NY, USA, 206–213. DOI:https://doi.org/10.1145/169059.169166
[63] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '90), Association for Computing Machinery, New York, NY, USA, 249–256. DOI:https://doi.org/10.1145/97243.97281
[64] Sergey I. Nikolenko, Sergei Koltcov, and Olessia Koltsova. 2017. Topic modelling for qualitative studies. *Journal of Information Science* 43, 1 (February 2017), 88–102. DOI:https://doi.org/10.1177/0165551515617393
[65] Mie Nørgaard and Kasper Hornbæk. 2006. What do usability evaluators do in practice? an explorative study of think-aloud testing. In *Proceedings of the 6th conference on Designing Interactive systems* (DIS '06), Association for Computing Machinery, New York, NY, USA, 209–218. DOI:https://doi.org/10.1145/1142405.1142439
[66] Freddy Paz, Freddy Paz, José Pow-Sang, and César Collazos. 2018. A Formal Protocol to Conduct Usability Heuristic Evaluations in the Context of the Software Development Process. *International Journal of Engineering and Technology(UAE)* 7, (May 2018), 10–19. DOI:https://doi.org/10.14419/ijet.v7i2.28.12874
[67] Freddy Paz, Daniela Villanueva, Cristian Rusu, Silvana Roncagliolo, and José Antonio Pow-Sang. 2013. Experimental Evaluation of Usability Heuristics. In *2013 10th International Conference on Information Technology: New Generations*, 119–126. DOI:https://doi.org/10.1109/ITNG.2013.23
[68] Elisa Maria Pivetta, Daniela Satomi Saito, Carla da Silva Flor, Vania Ribas Ulbricht, and Tarcísio Vanzin. 2014. Automated Accessibility Evaluation Software for Authenticated Environments. In *Universal Access in Human-Computer Interaction. Design for All and Accessibility Practice*, Springer International Publishing, Cham, 77–88. DOI:https://doi.org/10.1007/978-3-319-07509-9_8

[69] Daniela Quiñones, Cristian Rusu, and Silvana Roncagliolo. 2014. Redefining Usability Heuristics for Transactional Web Applications. In *2014 11th International Conference on Information Technology: New Generations*, 260–265. DOI:https://doi.org/10.1109/ITNG.2014.46

[70] Hootan Rashtian and Sathish Gopalakrishnan. 2019. Balancing Message Criticality and Timeliness in IoT Networks. *IEEE Access* 7, (2019), 145738–145745. DOI:https://doi.org/10.1109/ACCESS.2019.2944463

[71] Janet Read. 2015. Children as participants in design and evaluation. *interactions* 22, 2 (February 2015), 64–66. DOI:https://doi.org/10.1145/2735710

[72] Jeremy Rose and Christian Lennerholt. 2017. Low Cost Text Mining as a Strategy for Qualitative Researchers. *Electronic Journal of Business Research Methods* 15, 1 (2017), 2–16.

[73] Murat Şahin and Eren Aybek. 2019. Jamovi: An Easy to Use Statistical Software for the Social Scientists. *International Journal of Assessment Tools in Education* (December 2019), 670–692. DOI:https://doi.org/10.21449/ijate.661803

[74] Rodrigo Santos, Cláudia Werner, Heitor Costa, Ramon Abílio, and Hudson Borges. 2012. Managing reusable learning objects and experience reports in EduSE portal. In *2012 IEEE 13th International Conference on Information Reuse Integration (IRI)*, 631–638. DOI:https://doi.org/10.1109/IRI.2012.6303068

[75] Martin Schrepp, Manuel Pérez Cota, Ramiro Gonçalves, Andreas Hinderks, and Jörg Thomaschewski. 2017. Adaption of user experience questionnaires for different user groups. *Univ Access Inf Soc* 16, 3 (August 2017), 629–640. DOI:https://doi.org/10.1007/s10209-016-0485-9

[76] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2014. Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios. In *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience* (Lecture Notes in Computer Science), Springer International Publishing, Cham, 383–392. DOI:https://doi.org/10.1007/978-3-319-07668-3_37

[77] Aleksandra Slavkovic and Karen Cross. 1999. Novice heuristic evaluations of a complex interface. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '99), Association for Computing Machinery, New York, NY, USA, 304–305. DOI:https://doi.org/10.1145/632716.632902

[78] Ana Cecilia Ten and Freddy Paz. 2017. A Systematic Review of User Experience Evaluation Methods in Information Driven Websites. In *Design, User Experience, and Usability: Theory, Methodology, and Management*, Springer International Publishing, Cham, 492–506. DOI:https://doi.org/10.1007/978-3-319-58634-2_36

[79] M.J. Van den Haak, M.D.T de Jong, and P.J. Schellens. 2004. Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. *Interacting with Computers* 16, 6 (December 2004), 1153–1170. DOI:https://doi.org/10.1016/j.intcom.2004.07.007

[80] Verônica T. Vaz, Guilherme H. Travassos, and Tayana Conte. 2012. Empirical assessment of WDP tool: A tool to support web usability inspections. In *2012 XXXVIII Conferencia Latinoamericana En Informatica (CLEI)*, 1–9. DOI:https://doi.org/10.1109/CLEI.2012.6427210

[81] Viswanath Venkatesh, Susan A. Brown, and Hillol Bala. 2013. Bridging the Qualitative-Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems. *MIS Quarterly* 37, 1 (2013), 21–54.

[82] Sigurbjorg Groa Vilbergsdottir, Ebba Thora Hvannberg, and Effie Lai-Chong Law. 2014. Assessing the reliability, validity and acceptance of a classification scheme of usability problems (CUP). *Journal of Systems and Software* 87, (January 2014), 18–37. DOI:https://doi.org/10.1016/j.jss.2013.08.014

[83] Hubert Wagner, Paweł Dłotko, and Marian Mrozek. 2012. Computational Topology in Text Mining. In *Computational Topology in Image Context*, Massimo Ferri, Patrizio Frosini, Claudia Landi, Andrea Cerri and Barbara Di Fabio (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 68–78. DOI:https://doi.org/10.1007/978-3-642-30238-1_8

[84] Paweł Weichbroth. 2020. Usability of Mobile Applications: A Systematic Literature Study. *IEEE Access* 8, (2020), 55563–55577. DOI:https://doi.org/10.1109/ACCESS.2020.2981892

[85] Dennis Wixon. 2003. Evaluating usability methods: why the current literature fails the practitioner. *interactions* 10, 4 (July 2003), 28–34. DOI:https://doi.org/10.1145/838830.838870

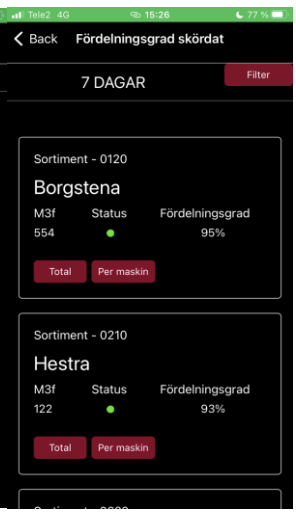
[86] Rosa Yáñez Gómez, Daniel Cascado Caballero, and José-Luis Sevillano. 2014. Heuristic Evaluation on Mobile Interfaces: A New Checklist. *The Scientific World Journal* 2014, (September 2014), e434326. DOI:https://doi.org/10.1155/2014/434326

[87] Chong Ho Yu, Angel Jannasch-Pennell, and Samuel DiGangi. 2011. Compatibility between Text Mining and Qualitative Research in the Perspectives of Grounded Theory, Content Analysis, and Reliability. *Qualitative Report* 16, 3 (May 2011), 730–744.
[88] Markos Zachariadis, Susan Scott, and Michael Barrett. 2013. Methodological Implications of Critical Realism for Mixed-Methods Research. *MIS Quarterly* 37, 3 (2013), 855–879.
[89] Panagiotis Zaharias and Panayiotis Koutsabasis. 2012. Heuristic evaluation of e-learning courses: a comparative analysis of two e-learning heuristic sets. *Campus-Wide Information Systems* 29, 1 (January 2012), 45–60. DOI:https://doi.org/10.1108/10650741211192046
[90] Eiman Zargaran, Nadine Schuurman, Andrew Nicol, Richard Matzopoulos, Jonathan Cinnamon, Tracey Taulu, Britta Ricker, David Brown, Pradeep Navsaria, and Sannia Hameed. 2014. The Electronic Trauma Health Record: Design and Usability of a Novel Tablet-Based Tool for Trauma Care and Injury Surveillance in Low Resource Settings. *Journal of the American College of Surgeons* 218, (January 2014), 41–50. DOI:https://doi.org/10.1016/j.jamcollsurg.2013.10.001
[91] Xiaofei Zhou, Yue Hu, and Li Guo. 2014. Text Categorization based on Clustering Feature Selection. *Procedia Computer Science* 31, (2014), 398–405. DOI:https://doi.org/10.1016/j.procs.2014.05.283
[92] Förarstöd krävs för att låsa upp skogsmaskinernas potential. Retrieved January 31, 2022 from https://www.skogforsk.se:443/nyheter/2018/forarstod-kravs-for-att-lasa-upp-skogsmaskinernas-potential/
[93] Jamovi. *Contingency Tables*. Retrieved May 4, 2022 from https://www.jamovi.org/jmv/conttables.html
[94] Topic Modelling. *Topic modelling: 2.2 Understanding the results*. Retrieved May 10, 2022 from https://port.sas.ac.uk/mod/book/view.php?id=620&chapterid=497
[95] Vida AB. *Vida AB*. Retrieved January 31, 2022 from https://www.vida.se/en/vida-ab/

APPENDICES

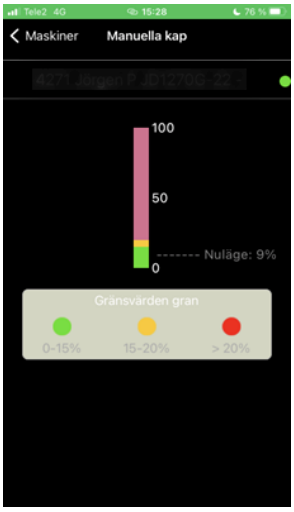
A1. Vida WQP Screenshots

1. Start page

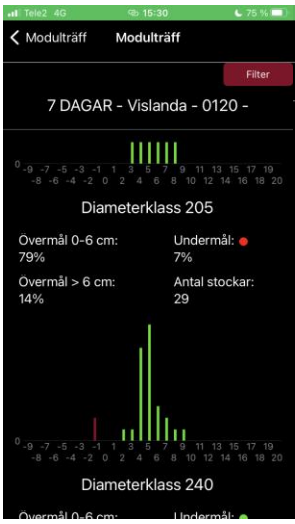


2. Sawmill- page

3. Filter



4. Manuella Kap



5.Diagram Modulträff

6. Vrak

7. Meny

A2. Heuristic principles for users (Swedish)

Id	Heuristic	Beskrivning	Exempel
1	Användandet av skärm utrymmet	Appens gränssnitt ska vara designat så att delarna av designen sitter på lagom avstånd från varandra. Inte för långt bort och inte för nära varandra.	Till exempel så är det viktigt att det inte finns för mycket saker på samma skärm, det riskerar att det blir rörigt och svårt att navigera vart man ska någonstans.
2	Följersamma standard	Appens komponenter bör sitta på samma ställe genom hela appen och bör följa samma designmönster. Detta för att stimulera användarens korttidsminne och undvika missförstånd	De knappar/funktioner som finns på flera ställen i appen , ska se likadana ut och fungera på samma sätt överallt. Till exempel knappen filter som i appen finns på flera ställen.
3	Synlig och enkel tillgång till all information	All information såsom text, bild, ljud och video som finns i appen ska vara synlig och läsbar. Det är viktigt att komponenterna på skärmen är korrekt justerade och designande så det står ut ordentligt.	Ett exempel på detta är att man i diagrammen som visas i appen ska kunna utläsa att data som den presenterar, inklusive vad som står på y och x- axeln.
4	Appens komponenter ska vara avsedda för dess funktionalitet	Användaren ska lätt kunna förstå vilken information som ska vara i en komponent. Användaren ska lätt förstå komponenternas egenskaper.	Till exempel så ska det vara lätt att förstå vilken information som presenteras i de funktioner som appen har.
5	Appens språk ska vara användarbaserat	Det är viktigt att appen använder sig av det språk som dess användare brukar. Det ska ske på ett naturligt sätt och inte kännas invasivt eller sätta användaren under någon form av stress.	Förstår man som användare alla ordval och beskrivningar som finns i appen. Till exempel att objekt betyder- markägare

6	Felhantering	Appen ska kunna förutse situationer som kan leda till fel genom användarinput. Om ett fel uppstår ska appen snabbt kunna meddela användaren att det har skett, och återgå till den senaste fungerande versionen. Om det är svårt att ordna så ska appen kunna ge användaren egna val till vart den vill gå.	Felhanteringen i Vidas app kan till exempel vara att ni som användare blir meddelade om mätningarna inte är korrekt uppdaterade.
7	Användarinput	Den data eller text som användaren lägger in i appen kan vara baserad på assisterande teknologi och kunna läggas in på ett praktiskt sätt. Appen ska alltid visa allt som användaren har lagt in så att användaren har full kontroll.	Till exempel, är det lätt att förstå hur ni kan filtrera informationen och att ni kan välja (det ni lägger in att filtrera på) på ett praktiskt sätt.
8	Enkel tillgång till Appens funktioner	Appens huvudfunktioner ska på ett enkelt sätt kunna lokaliseras av användaren, helst inte mer än två steg bort. De funktionaliteter som används mest kan det finnas genvägar till. Ingen funktionalitet i appen ska vara svår att hitta eller förstå.	Till exempel, är det lätt för er att komma åt det ni är mest intresserade av att se, är det någon funktionalitet ni anser är svårare att hitta till?
9	Omedelbar feedback	Appen ska omedelbart kunna ge användaren feedback om systemets status på ett enkelt och konkret sätt så att användaren förstår. Att kunna uppdatera något lokalt på appen istället för globalt är att föredra	Till exempel, är det lätt att förstå/se när siffrorna senast var uppdaterade? Går det att uppdatera funktionerna i appen? Alla tillsammans eller lokalt?

10	Hjälp och dokumentation	Appen ska ha en hjälp-knapp där vanliga fel och vilka lösningar det finns för felen listas.	Finns det en hjälpguide ni kan använda er av?
11	Minska användarens minnesbelastning	Användaren ska inte behöva komma ihåg saker som ligger på en annan sida än vad den är på. Informationen som finns på den sidan man är på ska innehålla all den informationen som krävs för att slutföra en uppgift	Behöver ni ta hjälp av ert minne hur man kommer till de olika funktionerna eller förstår man hur man ska göra för att se alla funktioner man vill titta på?

A3. HEURISTICS PRINCIPLES FOR EXPERTS (ENGLISH)

Id	Heuristic	Description	Example
1	The use of screen space	The app's interface should be designed so that the parts of the design are at a reasonable distance from each other. Not too far away and not too close to each other.	For example, it is important that there are not too many things on the same screen, there is a risk that it will be messy and difficult to navigate where you are going somewhere.
2	Consistency and standards	The app's components should be in the same place throughout the app and should follow the same design pattern. This is to stimulate the user's short-term memory and avoid misunderstandings	The buttons / functions that are found in several places in the app should look the same and work in the same way everywhere. For example, the filter button that is in the app in several places.

3	Visible and easy access to all information	All information such as text, image, sound and video contained in the app must be visible and readable. It is important that the components on the screen are correctly adjusted and designed so that it stands out properly.	An example of this is that in the diagrams shown in the app it should be possible to read the data it presents, including what is written on the y and x- axis.
4	Adequacy of the component to its functionality	able to easily understand what information should be in a component. The user should easily understand the properties of the components.	For example, it should be easy to understand what information is presented in the functions that the app has.
5	Adequacy of the message to he functionality and to the user	It is important that the app uses the language that its users use. It should be done in a natural way and not feel invasive or put the user under any kind of stress.	As a user, do you understand all the word choices and descriptions in the app. For example, do you understand that “objects” means “landowners”
6	Error prevention and rapid recovery to the last stable state	The app must be able to anticipate situations that can lead to errors through user input. If an error occurs, the app should be able to quickly notify the user that it has occurred, and return to the latest working version. If it is difficult to arrange, the app should be able to give the user their own choices for where they want to go.	The error handling in Vida's app can, for example, be that you as a user are notified if the measurements are not correctly updated.
7	Ease of input	input The data or text that the user enters into the app can be based on assistive technology and can be entered in a practical way. The app should always show everything that the user has entered so that the user has full control.	For example, it is easy to understand how you can filter the information and that you can choose (what you put in to filter) in a practical way.

8	Ease of access to all functionalities	The app's main functions should be easily located by the user, preferably no more than two steps away. The most commonly used functionalities can be shortcuts to. No functionality in the app should be difficult to find or understand.	For example, is it easy for you to access what you are most interested in seeing, is there any functionality you find more difficult to find?
9	Immediate and observable feedback	The app must be able to immediately give the user feedback on the system's status in a simple and concrete way so that the user understands. Being able to update something locally on the app instead of globally is preferable	. For example, is it easy to understand / see when the figures were last updated? Is it possible to update the functions in the app? All together or locally?
10	Help and documentation	The app should have a help button where common errors and what solutions there are for the errors are listed.	Is there a help guide you can use?
11	Reducing the user's memory load	The user should not have to remember things that are on a page other than what they are on. The information should contain all the information required to complete a task	Do you need the help of your memory to get to the different functions or do you understand how to use all the functions you want to use?

A4. Consent form

Participation in a heuristic evaluation of Vida WQP

Form of consent:

1. Background and aim	The study is carried out in collaboration with Uppsala University, Skogforsk and Vida AB.The aim with the study is to compare the results of evaluation performed by users as participants and an evaluation performed by experts as participants.
2. How will the study work?	The study will collect data through ten individual heuristic evaluations. The evaluations are expected to take approximately one hour. The data collected will be compiled with the data from other participants and the answers will not be traceable to specific identities. The data will be used for the purpose of the this study and for Skogforsk and Vida to gain knowledge of how feedback systems such as WQP are received by harvest operators (users).
3. What are the risks?	The researcher has stated that it is not possible to identify any risks for those who participate in the study.
4. Data management	The collected data will be handled by the researcher from Uppsala University, the collected data that is shared with SKogforsk and Vida has been anonymised. No recordings have been conducted during the evaluations, only notes have been taken.
5. Voluntary participation	Participation in the study is completely voluntary and can be interrupted at any time without further explanation.
6. Responsible	The study will be conducted by Madeleine Silverbratt with supervision from Franck Tétard, Uppsala University and Karin Ågren, Skogforsk

Informed consent

- I confirm that I have received this written and other oral information about the research study.
- I give my consent to participate in the study and know that my participation is completely voluntary.
- I am aware that I can terminate my participation at any time without explanation.
- I allow that the information I have received and that collected data about me is stored and handled electronically by study supervisors.

.....
Date

.....
Participants signature

.....
Name clarification

.....
Interviewer, [Interviewers name], Signature

A5. Questionnaire for gathering participants for the evaluation performed by users

Var med och utvärdera Vida WQP!

Hej! Mitt namn är Madeleine Silverbratt, jag studerar sista året på masterprogrammet människa-datorinteraktion och jag kommer under våren i samförstånd med Skogforsk att skriva mitt examensarbete inom ämnet användbarhet. Mitt arbete fokuserar på att utvärdera hur användarvänlig er app Vida WQP är. Med användarvänlig menas hur lätt Vida WQP är att förstå, navigera i och om den uppnår det syfte den är skapad för. För att ta reda på hur Vida WQP´s användbarhet är och på vilket sätt den skulle kunna förbättras så kommer jag hålla intervjuer och fokusgrupper med några av er som använder appen.

Jag skulle bli väldigt tacksam om ni ville svara på nio korta frågor, enkäten tar ca två minuter att fylla och det är bara jag som kommer se era personuppgifter, Vida och Skogforsk kommer få ta del av datan först när den har blivit anonymiserad. Dem av er som är intresserade av att vara med vid intervjuer eller gruppdiskussioner kan också ange era kontaktuppgifter, så hör jag av mig längre fram i vår. Att delta i undersökningen kommer innebära något/några intervjutillfällen alternativt något/några tillfällen i gruppdiskussion tillsammans med mig.

Tack på förhand och jag ser fram emot att träffa några av er under våren!

...

* Obligatoriskt

1. Namn *

Ange ditt svar

2. Ålder *

Ange ditt svar

3. Vilken region tillhör du? *

- ○ Öst
- ○ Väst
- ○ Syd

4. Hur många års erfarenhet av skördare och aptering har du? *

Ange ditt svar

5. Hur ofta använder du appen *

- ◯ Varje dag
- ◯ 2-3 ggr/vecka
- ◯ 1 gång/vecka
- ◯ Några gånger i månaden
- ◯ 1 gång i månaden
- ◯ Mindre än en gång i månaden
- ◯ Aldrig

6. På en skala 1-5 hur upplever du att Vida WQP fungerar. 1 inte bra, 5 mycket bra. *

1	2	3	4	5
◯	◯	◯	◯	◯

7. På en skala 1-5 skulle du vara intresserad att vara en del av undersökningen? 1 vill inte vara med, 5 vill vara med. *

1	2	3	4	5
◯	◯	◯	◯	◯

8. Telefonnummer

Lägg till ditt telefonnummer och/eller mejladress om du är intresserad att vara med i undersökningen så jag kan kontakta dig!

Ange ditt svar

9. Mejladress

Lägg till ditt telefonnummer och/eller mejladress om du är intresserad att vara med i undersökningen så jag kan kontakta dig!

Ange ditt svar

A6. Formal document All identified problems- User Evaluation

Link to User Evaluation - All identified problems

- User_evaluation_WQP.xlsx - All identified problems.pdf

A7. Formal document Unique problems identified- User evaluation

Link to User Evaluation - Unique identified problems

- User_evaluation_WQP.xlsx - Unique problems

A8. Formal document All identified problems- Expert evaluation

Link to Expert Evaluation - All identified problems

- Expert_evaluation_WQP.xlsx - All identified problems.pdf

A9. Formal document Unique identified problems- Expert Evaluation

Link to Expert Evaluation - Unique identified problems

- Expert_evaluation_WQP.xlsx - Unique problems

A10. Result of correlation test on User evaluation, UFrequency, USolution vale & URelevance

1. Correlation Frequency/Relevance

Contingency Tables

	URelevance (1-5)					
UFrequency	1	2	3	4	5	Total
1	3	5	6	3	8	25
2	1	3	4	2	0	10
3	0	2	1	1	2	6
4	0	1	1	0	0	2
5	0	0	1	2	1	4
Total	4	11	13	8	11	47

χ^2 Tests

	Value	df	p
χ^2	12.0	16	0.745
N	47		

2. Correlation Solution value/ Relevance

Contingency Tables

	URelevance (1-5)					
USolution value	1	2	3	4	5	Total
0.000	2	3	2	0	1	8
0.200	0	0	1	0	0	1
0.250	0	1	0	0	0	1
0.400	0	0	0	1	0	1
0.500	0	1	3	0	0	4
0.667	0	1	0	0	1	2
0.800	0	0	0	1	0	1
1.000	2	5	7	6	9	29
Total	4	11	13	8	11	47

χ^2 Tests

	Value	df	p
χ^2	29.8	28	0.374
N	47		

3. Correlation Solution value/Frequency

Model Fit Measures

Model	Deviance	AIC	R^2_{McF}
1	118	134	0.0149

Note. The dependent variable 'USolution value' has the following order: 0.000 | 0.200 | 0.250 | 0.400 | 0.500 | 0.667 | 0.800 | 1.000

Model Coefficients - USolution value

Predictor	Estimate	SE	Z	p
UFrequency	−0.282	0.211	−1.34	0.180

A11. Result of correlation test on Expert evaluation, EFrequency, ESolution vale & ERelevance

1. Correlation Frequency/Relevance

Contingency Tables

	ERelevance (1-5)					
EFrequency	1	2	3	4	5	Total
1	4	7	26	15	5	57
2	0	2	7	4	2	15
3	0	1	2	4	0	7
4	0	1	2	0	0	3
5	0	3	0	0	0	3
Total	4	14	37	23	7	85

χ^2 Tests

	Value	df	p
χ^2	23.6	16	0.099
N	85		

2. Correlation Solution value/ Relevance

Contingency Tables

	ERelevance (1-5)					
Esolution value	1	2	3	4	5	Total
0	1	2	6	7	1	17
1	3	12	31	16	6	68
Total	4	14	37	23	7	85

χ^2 Tests

	Value	df	p
χ^2	2.39	4	0.665
N	85		

3. Correlation Solution Value/Frequency

Model Fit Measures

Model	R	R^2
1	0.144	0.0208

Model Coefficients - Esolution value

Predictor	Estimate	SE	t	p
Intercept	0.7104	0.0804	8.84	< .001
EFrequency	0.0564	0.0426	1.33	0.188

A12. User word cloud

A13. Expert word cloud

A14. User topic-modelling

Topic Modelling

Latent Dirichlet Allocation

Number of topics: 10

Topics

1: add, menu, lengths, remove, places, frustrating, two, double, press, explain
2: via, dark, disappears, mode, ext, home, inconsistendtstandards, phone, different, measurement
3: know, length, user, entered, able, wrecks, feedback, machines, wrong, addition
4: versions, overload, keep, correct, important, outdated, necessary, old, possibility, spruce
5: bucking, frustration, parameters, light, wrecks, anything, frustrated, influence, addition, manual
6: functions, make, purpose, adding, decsriptions, fully, directly, sawmills, start, cuts
7: understand, purpose, difficult, lights, could, function, fully, decsriptions, color, blind
8: color, page, start, codes, something, blind, choose, delete, says, confusing
9: app, red, uses, settings, possible, error, handling, press, explain, double
10: information, good, missing, variables, volume, interesting, timber, data, see, enter

A15. Expert topic-modelling

Topic Modelling

Latent Dirichlet Allocation

Number of topics: 10

Topics

1: total, regarding, clarify, belongs, calculation, scroll, move, logical, consist, price
2: user, page, make, top, access, give, see, description, function, days
3: either, menu, functions, two, inställningar, logga, meny, mnu, icons, ut
4: information, add, space, remove, graph, difficult, easily, result, another, descriptive
5: text, screen, fit, instead, easily, misuse, choose, table, line, designed
6: displayed, two, screens, percentage, read, misslyckade, overview, lyckade, gran, separate
7: everything, buttons, userlanguage, correct, correctly, swedish, english, spelled, objects, sure
8: time, feels, periods, clear, different, view, money, interesting, informations, look
9: unclear, could, design, use, lot, know, numbers, something, percentage, total
10: headlines, headline, header, similar, designed, line, guide, app, visual, calculation