

Linda Salo

USABILITY AND COGNITIVE LOAD IN HEAVY MACHINERY

Master's Thesis
Faculty of Information Technology and Communication Sciences
Examiners: Päivi Majaranta
Toni Pakkanen
May 2025

ABSTRACT

Linda Salo: Usability and Cognitive Load in Heavy Machinery
Master's Thesis
Tampere University
Master's Programme in Computing Sciences
May 2025

Usability is an essential part of any product's user experience. The importance of usability is emphasized especially in the context of human-machine interfaces, which may inherently include complex functionalities. The amount of cognitive load is also known to increase in relation to the complexity of the product in question. When the product itself is intricate by nature, the inherent amount of cognitive load is higher when using the product.

This thesis examines the connection between usability and cognitive load in the context of human-machine interfaces of complex heavy machinery. The aim was to understand the relationship between usability and the cognitive load, and whether the amount of experienced cognitive load can indicate the presence of usability issues. Usability was evaluated with task-based usability tests while the cognitive load was measured with self-assessment methods during the evaluation sessions. The research included five usability tests with experienced end users.

The research concluded that usability and cognitive load did not seem to have obvious connection. While on some level a connection could be found between efficiency and cognitive load as well as satisfaction and cognitive load, the amount of experienced cognitive load alone did not indicate the presence of usability issues. However, the type of usability issues encountered seemed to impact the amount of experienced cognitive load. The findings suggest that even though a connection could be found in some aspects between the usability test results and cognitive load, they are not interchangeable metrics as they are measuring different constructs.

Keywords: Usability, cognitive load, usability evaluation, human-machine interface

The originality of this thesis has been checked using the Turnitin Originality Check service.

TIIVISTELMÄ

Linda Salo: Käytettävyys ja kognitiivinen kuormitus raskaiden työkoneiden käytössä
Pro gradu -tutkielma
Tampereen yliopisto
Tietojenkäsittelytieteiden tutkinto-ohjelma
Toukokuu 2025

Käytettävyys on keskeinen osa tuotteen käyttökokemusta. Käytettävyyden tärkeys korostuu erityisesti ihmisen ja koneen välisten käyttöliittymien kontekstissa, jotka voivat luonnostaan sisältää monimutkaisia toimintoja. Kognitiivisen kuormituksen määrän tiedetään myös lisääntyvän järjestelmän kompleksisuuden yhteydessä. Kun tuotteen käyttäminen itsessään on vaativaa, myös luontaisen kuormituksen määrä kasvaa.

Tässä pro gradu -tutkielmassa tutkittiin käytettävyyden ja kognitiivisen kuormituksen välistä suhdetta kompleksisten raskaiden työkoneiden käyttöliittymässä. Tavoitteena oli ymmärtää millainen yhteys käytettävyysongelmilla ja koetun kuormituksen määrällä on, ja indikoiko kognitiivisen kuormituksen määrä käytettävyyden ongelmia. Käytettävyyttä arvioitiin tehtäväperusteisella käytettävyydestillä, jonka yhteydessä käyttäjien kokemaa kognitiivista kuormitusta mitattiin itsearviointimenetelmän avulla. Tutkimuksessa tehtiin viisi käytettävyydestiä, joiden osallistajat olivat kokeneita loppukäyttäjiä.

Tutkimuksen tulosten perusteella merkittävää yhteyttä ei voida havaita käytettävyyden ja kognitiivisen kuormituksen välillä. Vaikka jollain tasolla yhteys löytyi tehokkuuden ja kognitiivisen kuormituksen välillä sekä tyytyväisyyden ja kognitiivisen kuormituksen välillä, käytettävyysongelmien esiintymistä käyttöliittymässä ei voida havaita tarkastelemalla pelkästään koetun kognitiivisen kuormituksen määrää. Sen sijaan vaikuttaa siltä, että käytettävyysongelmien tyypillä olisi yhteys koettuun kognitiiviseen kuormitukseen. Tulokset viittaavat siihen, että vaikka joiltain osin yhteys käytettävyydestien tulosten ja kognitiivisen kuormituksen välillä oli havaittavissa, ne eivät ole keskenään vaihtokelpoisia mittareita.

Avainsanat: Käytettävyys, kognitiivinen kuormitus, käytettävyyden arviointi, ihmisen ja koneen käyttöliittymä

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin Originality Check -ohjelmalla.

USE OF AI IN THESIS

I have utilised AI tools in my thesis:

- No
- Yes

I acknowledge that I am fully responsible for the entire content of my thesis, including the parts generated by AI, and accept accountability for any violations of ethical standards in publications.

PREFACE

This research was conducted for Metso, and I am thankful for the opportunity to study a topic that I personally found extremely interesting. I would also like to thank everyone at Metso who supported me during this journey.

I want to thank my supervisors Eeva-Maria Myllymaa at Metso and Päivi Majaranta at Tampere University for guiding me through the whole process and providing me invaluable feedback.

Finally, I am most grateful to my husband Atte and my son Elias for all the love and the laughs. I could not have done this without your support.

Tampere, 26 May 2025

Linda Salo

CONTENTS

1.INTRODUCTION	1
1.1 Motivation and Research Questions	2
1.2 Thesis Structure	3
2.USABILITY AND COGNITIVE LOAD	5
2.1 Usability and its Evaluation	5
2.2 Information Processing and Cognitive Load.....	8
2.3 Cognitive Load and Usability Evaluation	10
3.METHODS.....	13
3.1 Research Methods.....	13
3.2 Test Setting and Scope	14
3.3 Analysis Methods	16
4.RESULTS	18
4.1 Results Structure	18
4.2 Main Results.....	20
4.3 Detailed Results	25
5.DISCUSSION.....	35
5.1 Research Findings.....	35
5.2 Limitations and Future Research	38
5.3 Strengths and Contributions	39
6.CONCLUSIONS.....	41
REFERENCES.....	43
APPENDIX A: TRANSLATED NASA-TLX QUESTIONNAIRE.....	48

1. INTRODUCTION

Usability is essential in any product, and its importance is highly emphasized when developing complex automated systems for heavy machinery. When operating with human-machine interfaces the operators must be able to react quickly in diverse time-critical circumstances, such as possible malfunctions or alarm situations. Moreover, these machines are often operated in highly versatile, safety-critical environments. Therefore, it is necessary to make sure that the human-machine interface is easily understandable and straightforward to use.

In human-centered design process a central method for ensuring usability is to conduct effective user-centered usability evaluations. Usability evaluations can be conducted by utilizing different methods, from which the most appropriate ones should be selected. Various factors should be considered when selecting the most suitable one, such as the level of the design, stage of development process and cost effectiveness. (International Organization for Standardization, 2019) The earlier the usability evaluation has been adopted into the development process, the more inexpensive it also is to fix the potential issues (Salvador et al., 2014). However, if usability evaluations are ignored, on top of cost increases it also affects overall user satisfaction and performance. Furthermore, in the worst cases poor usability can expose the users to situations where their safety may be compromised. This further emphasizes the significance of usability evaluations in the field of heavy machinery. Moreover, it shows how the evaluation procedures should be integrated into the product development process from the beginning, as the level of usability can often be tested already with the first prototypes of the product with relatively inexpensive evaluation methods. (International Organization for Standardization, 2019; Norman, 2013)

One aspect to consider in relation to usability is to study the required cognitive load when using the product. When a user interacts with a product, the received information goes theoretically through different human memory systems, which have been conceptualized by various information processing models (Atkinson & Shiffrin, 1968/2024; Baddeley et al., 2012). Working memory is one of the central concepts, which handles all the information that is currently processed at a given time. While existing knowledge can be retrieved from the long-term memory and new information is first perceived through different senses, the working memory is considered as the central processor of all the

information. However, multiple studies have been conducted concluding that the working memory can process only a limited amount of information at a time (Cowan, 2001; Miller, 1956). When the amount of information exceeds the capacity of one's working memory, the excessive amount of cognitive load is causing temporary difficulties in the ability to process information or complete actions.

Cognitive load refers to the amount of effort that is needed from the user so that they are able to achieve their goals while using the product in question. Higher cognitive load means that more effort is required from the user, and conversely lower cognitive load indicates that not much effort is needed from the user. (Jaiswal et al., 2019) As human-machine interfaces are inherently cognitively challenging to use, particularly considering various additional factors such as safety and alarm situations that require immediate response from the operator, it is valuable to examine the required cognitive load especially in relation to usability.

Cognitive load theory divides the cognitive load further into three categories: intrinsic, extraneous and germane cognitive load (Sweller et al., 2011). Moreover, as Paas et al. (2003) concluded, the extraneous cognitive load is mostly caused by poor design choices, which demand the users to allocate even more mental resources into the activity in addition to what the main task inherently requires. Thus, one perspective in ensuring usability would be to focus on minimizing the required extraneous cognitive load (Sweller, 2010). Therefore, assessing the cognitive load in the usability evaluation process may provide valuable insights into the perceived amount of effort that the participants need to use during the tasks.

1.1 Motivation and Research Questions

Cognitive load has been studied in various complex and safety-critical contexts before, such as aviation, in-vehicle systems and health industry (Clarke et al., 2020; Pan et al., 2024; Ren et al., 2024). Especially in the case of human-machine interfaces, the cognitive load should be reduced to ensure as effortless operations as possible. As Khawaja et al. (2014) point out in their study, users operating complex, time-critical systems inherently experience higher mental effort, which can affect the level of performance in completing the tasks. Therefore, it can be inferred that if a task itself requires higher cognitive load, it is even more likely that the user will be exposed to erroneous actions if the product does not have good usability.

Even though usability and cognitive load of human-machine interfaces has been studied in various fields, to my knowledge no such studies have been conducted regarding

mobile rock crushers. Moreover, the connection between experienced cognitive effort and usability seems to lack information in the heavy machinery industry. This research aims to provide a clearer understanding of the topic by studying the subjective cognitive load and its relation to observed effort when using a new user interface designed for heavy machinery. On top of this, the aim is to gain insights on whether measuring subjective cognitive load alone can indicate challenges in the product's usability. Thus, the research questions of this thesis are:

1. How does cognitive load align with observed usability challenges during usability evaluation of a human-machine interface?
2. How well can measuring cognitive load indicate difficulties in usability?

For the purpose of finding answers to these research questions, five task-oriented usability tests were conducted with a high-fidelity Figma prototype of a new heavy machinery user interface to study the usability of the new design. The cognitive load was studied during the usability evaluation sessions with subjective measures in aims to measure the experienced cognitive load while trying to accomplish the given tasks. The scope of the tests included basic, daily operational tasks that the operators should be able to accomplish effortlessly. Moreover, all participants were experienced heavy machinery operators to ensure that the tasks at hand would not need additional learning but would be achievable based on previous knowledge of the topic.

1.2 Thesis Structure

In chapter 2 the theoretical background regarding usability, its evaluation, and information processing provides context for the research of this thesis. First, the concepts related to usability and its evaluation are introduced. After that, the chapter covers the theoretical concepts related to different information processing models to further provide understanding how the cognitive load is linked to memory and what are the limitations related to information processing and human memory. Finally, the connection between cognitive load and usability is presented in the chapter, while focusing on how cognitive load can be measured in relation to usability.

After that, the selected research methods for evaluating heavy machinery and measuring cognitive load are introduced in chapter 3. The chapter also covers the utilized analysis methods and procedures. Then, chapter 4 introduces the results of the study. First, the structure of the results is explained. After that, the main findings are introduced to provide an overall understanding of the results. Finally, the findings of the research are presented in more detail by one participant at a time. The results are discussed already in this

chapter at some level as the selected research methods have certain restrictions related to the subjectivity of self-assessment, which is why comparing pure results with each other would not be ideal.

In chapter 5 the study results are discussed in greater depth and from different viewpoints. The results are connected to previous research while pondering the implications of this study. Moreover, the selected methods are discussed concurrently. After that, the limitations of the study are addressed, and possible future research areas are introduced accordingly. In addition, the strengths and contributions of this research are examined on a more general level. Finally, in chapter 6, the conclusions from the research are presented.

2. USABILITY AND COGNITIVE LOAD

In this chapter the theoretical background of the research is introduced. First, usability and its evaluation are defined. Then, the concepts of information processing and cognitive load are presented, including the models and theories behind the concepts. Finally, the relation between cognitive load and usability in human-machine interaction design is discussed, and moreover the importance of studying cognitive load in terms of usability is argued.

2.1 Usability and its Evaluation

Usability is considered to be a part of the whole user experience when using a product (International Organization for Standardization, 2019). While user experience consists of more hedonistic concepts including the user's perceptions and anticipations towards a product, usability refers to the extent to which the product can be used (Hassenzahl & Tractinsky, 2006; International Organization for Standardization, 2019). Furthermore, usability is considered to consist of multiple different factors. Standards often take three aspects into account regarding the concept of usability, which are effectiveness, efficiency and satisfaction (International Society of Automation, 2015; International Organization for Standardization, 2018). In contrast, Nielsen's (1993) definition examines the construct by considering learnability, efficiency, memorability, errors, and satisfaction. However, regardless of the definition or dimensions of usability, it is considered to be a fundamental part of any product which should be carefully evaluated to provide the best possible user experience (Sharp et al., 2023).

In human-machine interfaces usability is extremely important due to the nature of the products. They may be inherently complex systems, while the context of use can include various situational factors which may affect the use of the interfaces. Moreover, operating human-machine interfaces may involve situations where the user needs to perform quick decisions while having physical constraints. (Palviainen et al., 2009) Furthermore, in some contexts the operators might be required to handle multiple situations at once, while simultaneously needing to interpret information from the interface in time-critical circumstances (Harvey et al., 2011). The importance of usability in human-machine interfaces has also been studied and emphasized in various fields. Barshi et al. (2024) examined two airplane crashes where poor usability of a human-machine interface was one of the central factors that led to the fatal accidents. According to their study, the visualization of the aircraft's user interface did not represent the actions in real life due

to the lack of symmetry in relation to how the pilot operated the aircraft. Similarly, Goel et al. (2017) studied alarm systems and alarm handling in human-machine interfaces, where they emphasized how multiple fatal accidents have occurred due to insufficient design of alarm systems. Thus, the level of usability has a critical impact to the overall safety and quality in human-machine interfaces, which may be compromised if usability is not ensured throughout the product development process. As poor usability may increase the risk of errors, impact the user performance and increase the level of frustration, the interface must be clearly understandable, so that the product itself does not cause any additional issues for the user (Hautamäki et al., 2017). Therefore, the usability of the product should be validated also considering the system's overall safety and reliability.

The ease of use can be ensured by inspecting the different dimensions of usability in relation to the product in question. This can be achieved by conducting usability evaluations throughout the whole product development process. (International Organization for Standardization, 2019) The types of usability evaluation methods can be categorized as usability inspection methods and user testing methods (Nielsen, 1993). When evaluating a product through usability inspection, such as assessing the product against heuristics, the evaluation process is often conducted by usability experts. However, all inspection methods do not involve actual users in the evaluation procedure but are mainly based on the inspector's expertise and assumptions. (Sharp et al., 2023) In contrast, user testing methods are usually conducted with real users, which are beneficial to gain understanding of the users, their behavior, and to acquire deeper insights about the use of the product in question. (Nielsen, 1993)

One of the most commonly used user testing method is a task-based usability test (Paz & Pow-Sang, 2014). During a usability test session, the recruited participant would try to accomplish pre-defined tasks, which would be given to them one at a time. To understand the participant's thought processes, think-aloud method is often utilized during these test sessions, meaning that the participants would say out loud all their thoughts that may arise while they are attempting to accomplish the tasks. (Sharp et al., 2023) However, this is not always uncomplicated and might skew the data that would be examined in the analysis process (Hertzum et al., 2009). For example, if efficiency is studied by inspecting task times, the think-aloud method might cause the time on task to increase due to the unfamiliarity of the method. Furthermore, if a participant is particularly talkative, they may talk rather long times, which may unintentionally increase the task times. As these different evaluation methods have their specific strengths and weaknesses, according to Tan et al. (2009) heuristic evaluation and traditional usability testing

methods are complementary to each other and should be utilized respectively throughout the whole product development process. In their study it is recommended to conduct heuristic analysis in the earlier stages, while the usability testing would be beneficial in the later stages of the development process. Combining different evaluation methods would allow the most critical usability issues to get recognized in the earlier stages, when fixes would not consume the resources as much as if the issues would be found after the product has already been implemented. While there are many different methods available for evaluating a product's usability, according to Harvey et al. (2011) the most suitable methods should be selected by considering the available resources, the goal of the evaluation, the stage of the product development and the people required for the evaluation process. Nevertheless, whatever methods may be chosen for evaluating a product's usability, the outcome of the usability evaluations is to understand how easy the product is to use considering the end user's point of view.

When selecting the most suitable usability evaluation method, the type of collected data should also be considered. For example, while quantitative metrics can reveal how many issues may emerge, or how many users encounter the same specific issues when using the interface, the numbers alone will not give any explanation considering the reasons why users may struggle while using the product. This is why gathering also qualitative data is important. When doing qualitative usability testing it is possible to dive deep into the actual problems and get more detailed information about the different aspects of possible issues. (Sharp et al., 2023) Therefore, data triangulation should be included in the usability evaluation methods to ensure that different viewpoints are taken into consideration. Triangulation refers to collecting various types of data, which is considered to validate and strengthen the results of the evaluation. In a study by Oliver et al. (2010) data triangulation and its importance in user evaluation were discussed, and its importance was presented through the results, especially regarding user satisfaction. During usability testing, this can be achieved for example by observing how the users interact with the system, quantifying the collected data and collecting qualitative insights to discover the users' perceptions, including in-depth understanding of the main issues in the use of the system.

One important aspect to consider while evaluating a product's usability is the context of use. This includes for example the environment where the product is likely used, and how various environmental factors may affect the ease of use. (International Organization for Standardization, 2019) For example, possible issues regarding the context of use may not be so apparent in the development stage. In contrast, when going to the actual

use context the results might differ a lot and new issues may emerge due to the situational influence.

2.2 Information Processing and Cognitive Load

The human information processing studies can be dated to the middle of 1900s when cognitive psychology started to take more space from behaviorism, which was the reigning theory in the field of psychology from 1920 to 1950 (Mandler, 2002). While behaviorism aimed to study humans through their actions and the environment, the approach was not addressing the central issues of the mental processes behind the actions (Watrin & Darwich, 2012). Therefore, in 1960s the study of cognition and mental processes started to take more prominence over behaviorism (Mandler, 2002), and one of the influential factors related to cognitive psychology was the advancements in the field of computer science. The emergence of this new field provided a new analogy for psychologists to approach the human information processing in the same way as the information processing was functioning in computers (Baddeley, 1997).

The concept of human memory is often divided into different types of memory. One of the first information processing models were proposed by Atkinson & Shiffrin (1968/2024), which separates the memory system into three different stages: sensory register, short-term memory, and long-term memory as seen in Figure 1. First, the sensory register processes the quick, momentary acts where information received via a sense, such as touch or vision, is transmitted to one's brain. Next, the short-term memory refers to a system where limited amounts of information are being stored for a short period of time. In contrast, the long-term memory can be considered as a more permanent storage where the perceived information is transferred from the short-term memory by rehearsal and stored for long periods of time. In this model the information processing is conceptualized as a linear, systematic flow starting from perceiving the information to storing it to the long-term memory.

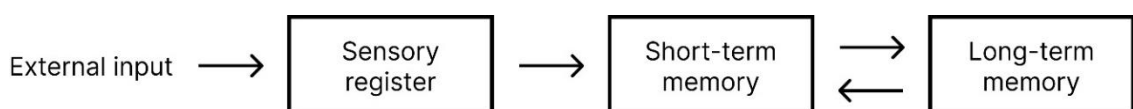


Figure 1. Information processing model according to Atkinson & Shiffrin

However, multiple versions of the human memory system and information processing models have been proposed since the Atkinson and Shiffrin's model was introduced in 1968, to provide more descriptive and comprehensive understanding of these concepts

(Baddeley, 2007; Cowan, 1998). One of the most widely accepted adaptation is the Baddeley's model, which is considered as an alternative approach to information processing. As shown in Figure 2, the Baddeley's model introduced a new component, working memory, to the information processing model, which is an additional concept related to the short-term memory (Purves et al., 2013).

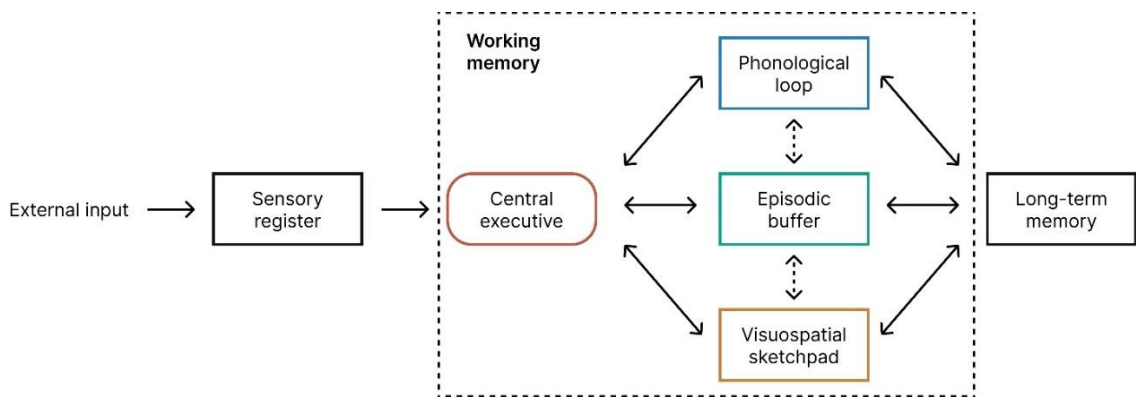


Figure 2. Information processing according to Baddeley's model

While the short-term memory is able to recall and retrieve information, the working memory is a representation of a temporary space for executive functions, such as problem-solving (Baddeley, 2007; Engle et al., 1999). Furthermore, the model separates areas for different functions such as processing language or visual input. Baddeley et al. (2020) believe that the information processing flow cannot be described as simply as in the model by Atkinson and Shiffrin (1968/2024), as its linear form does not consider the relation between the different memory types. Moreover, multiple studies have proven that the long-term memory impacts on how attention is focused (Moore et al., 2003; Olivers, 2011; Zimmermann et al., 2017). Thus, in contrast to the more linear approach it can be concluded that the received input to sensory register is impacted by one's ability to be aware of their surroundings in the first place. Furthermore, this means that the linear continuum starting from the sensory input and ending to long-term memory is not a sufficient representation of the human information processing.

In addition to modelling the information flow, studies have shown that there are limits in the amount of information that can be processed at a time. According to Miller's (1956) information processing theory, all information in the short-term memory is processed as chunks, and the capacity of holding information in short-term memory is limited. In their study, it was concluded that the short-term memory would be able to hold only approximately 7 chunks of information at a time. However, more current studies on the

limitations of short-term memory are suggesting that the previously believed approximation of the number of chunks might not hold true, and the capacity might be even more limited than what Miller had presented, reducing the number of chunks to 4 at a time (Cowan, 2001). Nevertheless, the common consensus is that the capacity of information processing in short-term memory is limited, regardless of the theorized number. When this capacity is reaching its limits, the phenomenon is described as a cognitive overload.

Cognitive load refers to the amount of effort that is needed for a person to perform an action. Cognitive load theory was introduced to examine how cognitive load affects people's information processing and problem-solving skills (Sweller, 1988). The theoretical framework divides cognitive load into three different categories: intrinsic, germane, and extraneous cognitive load (Plass et al., 2010). The intrinsic cognitive load refers to the inherent level of difficulty or complexity regarding the task at hand. The germane load considers the effort that is used for the learning activity itself, such as forming new mental models. Finally, the extraneous cognitive load addresses the difficulty of unnecessary effort that is required and is mainly caused by poor design or the amount of unnecessary information that is not needed. It has been studied how the excessive amount of cognitive load affects the performance and usability of systems (Clarke et al., 2020; Ren et al., 2024). Especially in the context of human-machine interfaces, where the intrinsic cognitive load is often high due to the complexity of the system itself and the challenging environments regarding the context of use, the impact of cognitive load should be noted when developing such products.

2.3 Cognitive Load and Usability Evaluation

As previously concluded, high cognitive load affects both a person's ability to perform desired actions and their overall performance during the activity. Furthermore, the amount of required effort during the use of a product affects the whole user experience. Therefore, high cognitive load could be an indication of low level of usability. Sweller (1994) suggests that instructional design may reduce the extraneous cognitive load, as the main reason for the extraneous type of cognitive load is considered to be a result of poor design choices. Thus, there is a natural connection between cognitive load and usability. Moreover, considering the required effort in addition to the ease of use when evaluating a product's usability may reveal additional insights from the user's point of view, which otherwise may not be observable.

In the context of human-machine interfaces, it is necessary to acknowledge the intricacy of the systems. As concluded before, it should be considered how the system's inherent complexity may affect the ease of use by increasing the intrinsic cognitive load.

Furthermore, the varying context of use should be considered even more consciously, as it can provoke straining situations for the operator. For example, alarm handling is considered as one of the most safety critical areas in the use of human-machine interfaces, which are considered as inherently strenuous for the operators (Goel et al., 2017). However, according to Nachreiner et al. (2006) the principles related to human factors are often neglected especially in the alarm handling situations, which cause the operators to experience unnecessary strain in already stressful circumstances. Therefore, it is crucial that these inherently straining situations are not further affected by poor usability, as it may further increase the experienced cognitive load of the systems. Thus, the usability of human-machine interfaces should be evaluated carefully also from the perspective of required mental effort and ensure that the operators are not exposed to potentially harmful situations.

Different kinds of cognitive load measurements have been utilized in usability studies to assess the experienced cognitive load, which include both objective and subjective methods. Objective methods incorporate different physiological assessments, which are measured at the same time as a participant executes an activity. Collet et al. (2014) studied the effects of various braking actions while driving, by measuring the participants' electrodermal activity to indicate the stress levels during the braking event. Moreover, various eye-tracking measures have been successfully utilized to study cognitive load in nuclear power plants by Gao et al. (2013), where they examined the pupil size, blink frequency and blink duration during emergency operation procedures. While these methods provide accurate results, they require external devices from the evaluator which are not always available.

In contrast, subjective methods for measuring cognitive load include observing the participants' behavior while they perform actions or asking the participants to self-evaluate the amount of effort they experienced during an activity. These are more accessible methods, as simple questionnaires with pen and paper are needed to conduct the study. One of the most common subjective methods includes the NASA Task Load Index (NASA-TLX) questionnaire developed by Hart & Staveland (1988) for National Aeronautics and Space Administration to study the mental workload during complex operations (Frazier et al., 2022).

The NASA-TLX questionnaire includes six different subscales, measuring different dimensions of effort during a task on a scale of 0 to 100. The subscales include mental demand, physical demand, temporal demand, performance, effort and frustration. After each task, the user is asked to assess the importance of each subscale by doing comparative evaluation with all subscales. However, the cognitive load can also be assessed

without weighting the subscales, when the method is often referred to as Raw NASA-TLX. In this method, the average scores from each questionnaire are calculated, and no weighting is added to the subscales. The validity of the NASA-TLX questionnaire without weighting the subscales has been studied in various fields, concluding that there is no significant difference whether the weighting method is utilized or not (DiDomenico & Nussbaum, 2008; Said et al., 2020). Furthermore, excluding the pairwise comparison streamlines the self-assessment process especially when it is used in usability testing, and the participants are asked to evaluate their experienced cognitive load multiple times in each test session.

While the cognitive load assessment with the NASA-TLX questionnaire has been proven to provide reliable results, contradictory findings have been found generally regarding the self-evaluation methods and the psychological bias related to such methods. According to Deffuant et al. (2024) people tend to overestimate their capabilities when asked to conduct self-assessments. However, research conducted by Gadsby & Hohwy (2023) suggested that an imposter phenomenon may conversely result in more negative self-evaluations. This indicates that the subjective method for measuring cognitive load may not be unambiguous, but there might be biases present in both positive and negative directions which should be acknowledged.

3. METHODS

This chapter covers the selected research methods that were utilized to study the cognitive load and evaluate usability. First, the research process and methods are explained, including what type of data is collected and how the insights and metrics were gathered. Next, the test setting is outlined, where the practicalities are described including the whole testing procedure. Finally, the selected methods for analyzing the results are discussed.

3.1 Research Methods

The product under evaluation in this study involved a complex automated system of a mobile rock crusher, which has an interface where both physical controls and a touch-based digital display are combined. However, in this research the focus was only on evaluating the usability of the digital display. Moreover, at the time of the research, the product was still under development. Therefore, due to confidentiality reasons, no images could be shown and no detailed information about the user interface could be explained in the results, and these restrictions affected the research study design and which analysis methods were selected.

As discussed in the previous chapter, there are many different usability methods that can be utilized to gain insight into a product's usability (Paz & Pow-Sang, 2014). However, as the aim of this research was to observe the possible usability challenges in combination with studying the users' subjective cognitive load, the research problems were approached with qualitative usability evaluation methods to be able to study the behavioral factors during the use of the system. Furthermore, with this evaluation method the analysis can be done at a higher level, without needing to show the user interface. The usability evaluation was organized by conducting task-oriented usability tests while utilizing the think-aloud method.

One of the main objectives in this research was to study each participant's cognitive load while using the new interface. While there are different subjective and objective methods that can measure the cognitive load, the subjective methods are easier to administrate and do not require any additional equipment. Therefore, the experienced cognitive load was measured with a subjective method, utilizing a Finnish translation of the NASA-TLX questionnaire developed by Hart & Staveland (1988), which can be found in Appendix A. Furthermore, self-evaluation provides insights into satisfaction, which is considered

as one dimension in the definition of usability (International Organization for Standardization, 2018; Nielsen, 1993).

Originally, the NASA-TLX questionnaire was developed for studying the perceived cognitive load during complex operational tasks when operating human-machine interfaces. The questionnaire has been validated in many studies from multiple different fields and yielding successful and reliable results (Clarke et al., 2020; Pan et al., 2024). Therefore, it would be a suitable method also for the purpose of this study which involved a complex human-machine interface.

The NASA-TLX questionnaire is used as a post-task questionnaire for assessing the subjective cognitive load after an activity. The original questionnaire includes 6 subscales: mental demand, physical demand, temporal demand, performance, effort and frustration. In practice, the participant fills out each subscale after an activity to assess how much effort they experienced during an activity, considering these different viewpoints. Furthermore, the original cognitive load measurement includes a weighting method where the participant makes pairwise comparisons with the subscales. In the weighting part the participant would compare two subscales with each other at a time, so that all subscales would be compared with each other. However, the weighting method adds a lot of additional burden to the participant and makes the test sessions heavier, especially when it is used in task-based usability evaluations where many tasks are often included in one test session. Thus, to measure the cognitive load during the usability evaluation, the participant was only asked to fill out the questionnaire after each task, excluding the pairwise comparisons.

3.2 Test Setting and Scope

In total, five real end users were recruited to participate in this study. The most important criterion for participation was to represent either one of the main end user personas, operator or site supervisor. These user personas have comprehensive knowledge of the practical work in the field, and they operate with heavy machinery on a regular basis. Thus, the selected personas represent the wanted user group. This specific criterion was set, because understanding the real users' thoughts and behavior is crucial for product development in terms of usability. The scope of this usability evaluation included only the main operational functionalities, considering the daily usage of the product. Moreover, as concluded before, alarm handling is considered one of the most critical tasks in a human-machine interface. Therefore, one of the tasks involves handling an alarm situation.

The usability evaluation was conducted with a high-fidelity Figma prototype of the designed interface. The prototype was accessed with a tablet, which was placed in front of them on a stand. A tablet was chosen as the device for testing interface as it provides the closest resemblance to the digital display of the real machine. Both the tablet and the real display are operated with touch-based interactions and are also quite the same in size.

In the beginning of each test session the participant was asked to sign a data privacy notice, where details about the collected data and its handling and storing were explicitly explained to the participants. It was also clarified that they have the right to withdraw their participation at any point. The instructions included an explanation about the purpose of the study, specifying that the test session is conducted to gain understanding about the system and evaluate its ease of use. It was emphasized that the focus is on testing the user interface, and not about testing the participant or their skills.

The participants were informed how the test procedure goes overall, and that the tasks are going to be revealed to them one at a time. Each test session consisted of five different tasks, which the participant tried to accomplish, which can be seen in Table 1. These tasks were given to the participant one at a time, where the scenario was outlined to form a mental picture of the situation when using the system. The participants were instructed how to fill out the NASA-TLX questionnaire. It was said that they should give ratings based on the latest task.

Table 1. *Task topics in the usability test sessions*

Task order	Task topic
Task 1	Find information
Task 2	Start crushing process
Task 3	Decrease feeder speed
Task 4	Change crusher setting
Task 5	Handle alarms

The participants were instructed to think aloud during the tasks to gain deeper insights about their thought processes, and to reveal possible challenges while using the system. It was encouraged to talk all thoughts out loud while being honest and without filtering their thoughts. Moreover, it was instructed that after the participant feels that they have

accomplished a task or if they want to move on to the next one, they should explicitly say it out loud.

3.3 Analysis Methods

As one of the main goals was to study the observed effort in usability evaluation, the observed usability issues were analyzed by gathering them into rational themes. Furthermore, to study effectiveness and efficiency of the new interface design the task success and number of clicks are examined in the analysis on each task. Finally, the cognitive load scores from the NASA-TLX questionnaires contribute to the third usability dimension of satisfaction. The average scores for each task were calculated from the participants' assessments. The results were examined one test session at a time, task by task. First, the observed usability issues were analyzed while counting the number of clicks. Next, the cognitive load scores were calculated and combined with the usability analysis. The specific order was important to avoid possible bias when observing the usability challenges.

The results from the task-oriented usability tests were analyzed by first watching through the videos from all usability test sessions. After the data had been reviewed, the videos were watched again, while concurrently documenting the encountered usability issues. These issues were then organized and grouped into higher-level themes, such as challenges related to understanding different button states or navigational issues related to findability.

Measuring the number of clicks adds quantifiable metrics to the usability analysis by revealing the efficiency of the use. While one of the most common measurements for efficiency is to examine the task times, it was not seen as the most suitable method in this research. When think-aloud method was utilized during the tasks, it inflicted longer narrations in some of the test sessions including lengthy reflections about their previous experiences of working in the field. Therefore, the task times would be biased and causing them to not reflect the actual time that was used during the activities. Thus, to provide stronger credibility of the research and triangulate the results, efficiency was measured by calculating the number of clicks. Moreover, task successes were examined to understand the level of effectiveness of which the participants were able to accomplish the tasks.

As previously explained, the research utilized the Raw NASA-TLX scores to assess the experienced subjective cognitive load from each task. The average scores result in a number between 0 and 100, where higher number indicates higher experienced cognitive

load during a task. When analyzing the results, it is noteworthy to acknowledge that people might use the scale in very different ways. For example, some may be more optimistic and careless by nature, and give lower overall scores than others. Therefore, the results from the NASA-TLX questionnaire were first analyzed by one participant at a time. Moreover, one participant's answers were examined in relation to all of their answers, before inspecting the possible trends between among all participants. Furthermore, it should be noted that the average cognitive load results should be considered by examining whether a participant gave higher cognitive load scores when compared to other tasks. This further supports the approach of analyzing the results of the research one participant at a time.

4. RESULTS

This chapter covers the results of the research. First, the structuring of the results is presented to guide the interpretation of the findings. The themes that emerged from the usability evaluation sessions during the analysis are introduced, and each theme is explained in more detail to provide understanding about what kind of challenges the themes cover. The subjective cognitive load ratings are explained, as well as how they should be assessed. The second part of this chapter addresses the main findings of the research, while discussing the differences and similarities that emerged regarding the subjective cognitive load and usability challenges. Finally, the results from each usability test session are explained in detail by one participant at a time.

4.1 Results Structure

First, the main results are presented to give an overview of the findings. After that, the detailed findings from the usability evaluation sessions will be presented and structured individually by one participant at a time. The participants are abbreviated as P1, P2 ... P5 when needed. Similarly, the tasks are abbreviated as T1, T2 ... T5 when addressing them in the text and in the tables.

The themes are based on the findings from the usability issue analysis. The main issues that emerged during the analysis process are included in the results tables. In total, there are nine different themes of usability issues which were repeated throughout the five different test sessions:

1. Button – interaction
 - Difficulties understanding the interactions of buttons
2. Button – state
 - Difficulties understanding the state that a button was trying to convey, for example whether the crushing process or devices such as feeder was on or off
3. Distinguishing non-interactive elements
 - Difficulties understanding which elements on the interface could be clicked and which were non-clickable

4. Navigation – findability
 - Difficulties navigating in the user interface
5. Navigation – interaction
 - Difficulties understanding how to interact with the navigation
6. Notification – status
 - Difficulties differentiating notification statuses, such as active or inactive alarms
7. Notification – type
 - Difficulties differentiating the notification types, such as alarms or warnings
8. Prototype issue
 - Occurrences where the prototype itself caused confusion due to lack of real-time feedback that would have been expected from the actual machine
9. Understanding feature functionalities
 - Situations where specific larger functionalities were causing confusion

The results will include the average Raw NASA-TLX score for each task. The average score will be in the range from 0 to 100, where a higher number indicates a higher experienced cognitive load. However, it is notable that there can be a lot of variation in the subjective measures and specifically in how each participant is scaling their answers. Therefore, the scores are also assessed by each participant at a time so that their overall scaling of the self-assessment can be examined while inspecting the connection between the observed usability issues and the number of clicks in the tasks.

The number of clicks is included in the results tables. In addition, the number of required clicks is included to provide understanding on what is the minimum number of clicks the participant needs to do during a task to help interpret the efficiency. In the individual analysis tables the number of clicks in each task in the test is titled as “Test” and the needed number of clicks is titled as “Required”.

Furthermore, there were larger themes present in some of the tests. For example, multiple participants reported that they do not understand English, which was the language used in the prototype interface. Thus, it can be assumed that the language caused

difficulties in understanding the information throughout the test sessions. Each test session where a participant reported the interface's language as an issue is discussed in the results section but is not included in the table of findings as the language issue can be considered to affect all tasks.

When interpreting the results, it should be noted that the inherent difficulty of the activities varies by task. While T1 may be considered as the easiest out of all tasks, focusing on finding correct information from the interface, tasks 2 to 4 are a bit more demanding as they require the participant to not only find the correct place but execute actions. The last task, T5, can be considered the most strenuous and stressful by nature, as it requires the participant to understand information and handle alarm situations.

4.2 Main Results

When examining the experienced cognitive load scores from the NASA-TLX questionnaires between the participants, the variation in the individual scaling of responses became prominent as can be seen in Figure 3. The biggest outlier in the data was found from the answers of P5 who reported very low cognitive load scores throughout the test session. In contrast, P3 gave quite high overall scores during their test session. However, P3 did seem to be most visibly frustrated and stressed during the whole test session which may expectedly affect their experienced effort, while P5 seemed to be quite care-less and calm during the test. Moreover, even though P3 reported that they experienced quite high cognitive load throughout the whole evaluation session, a clear decrease can be seen in the experienced cognitive load in T4 in which they did not have any observable usability issues and which they efficiently completed. This indicates that their answers are nevertheless aligned throughout the test results. However, these variations in the experienced cognitive load between the participants may suggest that their attitudes and current mental state might affect how they are scaling the answers. Moreover, it is important to acknowledge the subjectivity of the assessment, meaning that there might be biases in the self-evaluation of the experienced effort in both positive and negative directions as discussed previously.

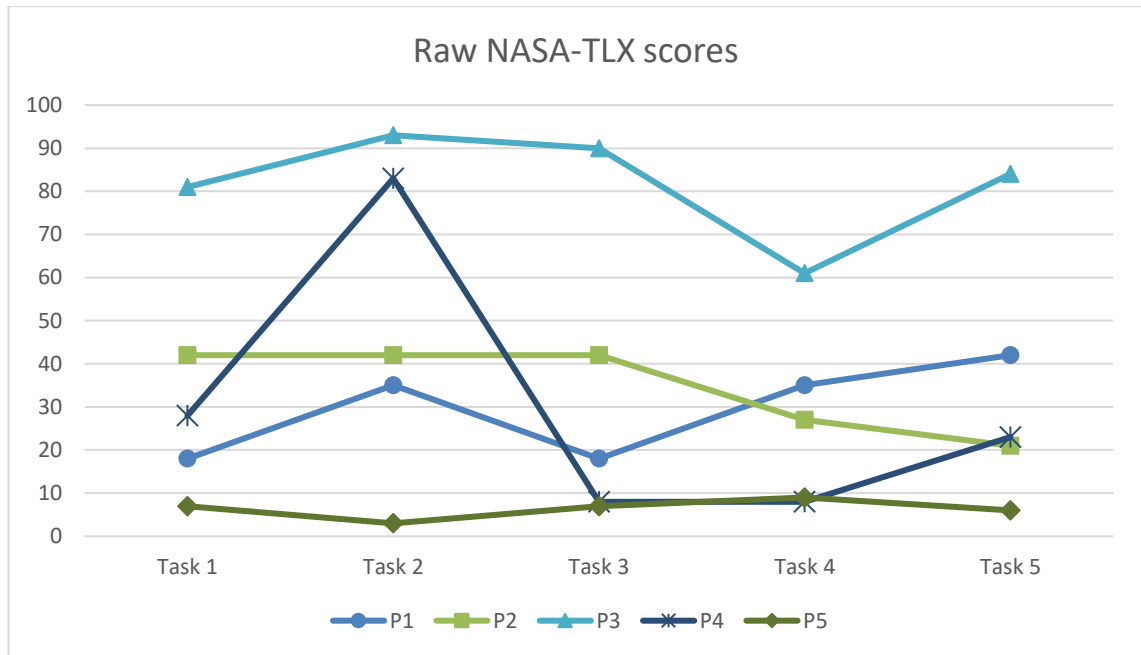


Figure 3. Raw NASA-TLX scores from each participant per task

When analyzing the connection between usability and cognitive load, there are different viewpoints which may explain how they are interconnected. The findings suggest that there is a connection between the experienced cognitive load and efficiency in most of the tasks as can be seen in Figure 4. If the efficiency is considered to be poor when a participant has clicked over double the number of clicks in relation to what is required to accomplish the task objective, a clear connection between the experienced cognitive load and efficiency can be seen in most of the tasks. When there are a bigger number of clicks during a task indicating poor efficiency, the experienced cognitive load is concurrently higher in most occurrences. In contrast, if a participant has accomplished the task efficiently, their experienced cognitive load is usually lower. However, in some tasks the connection is not obvious. For example, when examining P4 in T4 the number of clicks is 16 out of the required 3 clicks to complete the task and the cognitive load score is only 8, being the lowest that they reportedly experienced during the test session. Furthermore, as P5 gave consistently low cognitive load scores in every task, their results do not show any connection regarding efficiency. When examining the cognitive load of P5 in T5, they gave a cognitive load score of 6 even though they were visibly confused during the task and clicked 40 times out of the required 4 clicks.

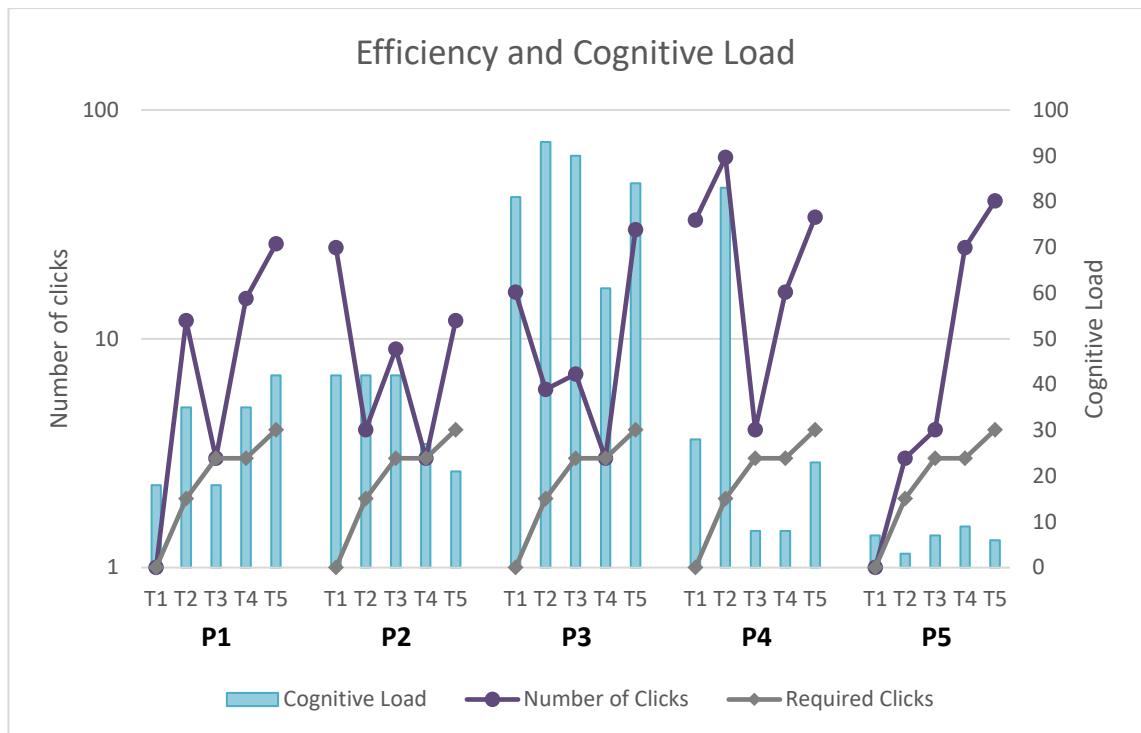


Figure 4. Number of clicks and the cognitive load scores in all tasks

Another viewpoint considers the amount of different usability issues in relation to experienced cognitive load. The results indicate that the number of issues alone did not seem to strongly affect the experienced cognitive load, as can be interpreted from Figure 5. For example, P3 had three different types of usability challenges in T5, but the experienced cognitive load was rated only as the third highest when compared to the scores in their other tasks. In contrast, P3 experienced the highest cognitive load in T2 where they had only two different usability problems. Similar connections could also be seen in other tests. For example, P2 had at most two different usability issues in tasks 1 to 3 in which they experienced the highest cognitive load. Yet, in T5 they had four different usability issues in total but experienced the lowest cognitive load. Thus, the number of usability issues may not fully reflect the amount of experienced cognitive load even when some relation was visible in a few tasks. Furthermore, these results suggest that the type of usability issue is more related to the experienced cognitive load.

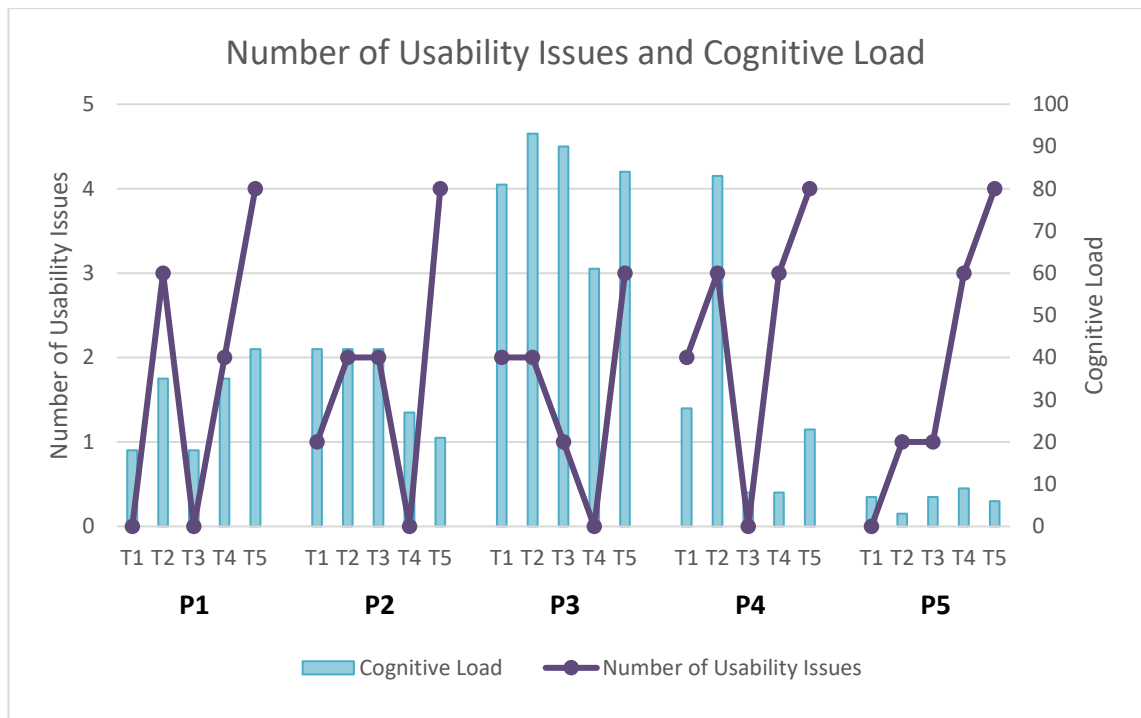


Figure 5. Number of found usability issues in relation to cognitive load

When examining the different types of usability issues and the Raw NASA-TLX scores, the results may indicate that the type of issues affect the experienced cognitive load. For example, in T1 the most frequent issues considered navigating in the system, which P2, P3, and P4 encountered during the task as can be seen in Table 2. Moreover, they also experienced the highest cognitive load in that task compared to other participants. However, in T5 the same participants, P2, P3 and P4, experienced a bigger number of different issues regarding understanding different states and interactions, yet their experienced cognitive load was staying approximately on the same level as in T1 or was reported to be even lower. This would suggest that the issues with navigating in the system were causing more cognitive strain than not understanding the states or interactions as in T5. Furthermore, the findings imply that there may be a stronger connection between the type of usability issues and the experienced cognitive load compared to the mere number of different usability issues and experienced effort.

Table 2. Findings from all usability test sessions

	Usability issues, Raw NASA-TLX & efficiency	P1	P2	P3	P4	P5	AVG (Eff %)
T1	Button – interaction			X			
	Navigation – findability		X	X	X		
	Navigation – interaction				X		
	Raw NASA-TLX	18	42	81	28	7	35
	# of clicks (required: 1)	1	25	16	33	1	15 (6.6 %)
T2	Button – interaction			X	X	X	
	Button – state		X	X	X		
	Distinguishing non-interactive elements	X			X		
	Navigation – findability	X	X				
	Navigation – interaction	X					
	Raw NASA-TLX	35	42	93	83	3	51
	# of clicks (required: 2)	12	4	6	62	3	17 (11.7 %)
T3	Button – state					X	
	Navigation – findability		X				
	Understanding feature functionalities		X	X			
	Raw NASA-TLX	18	42	90	8	7	33
	# of clicks (required: 3)	3	9	7	4	4	5 (60 %)
T4	Button – interaction				X	X	
	Prototype issue	X			X	X	
	Understanding feature functionalities	X			X	X	
	Raw NASA-TLX	35	27	61	8	9	28
	# of clicks (required: 3)	15	3	3	16	25	12 (25 %)
T5	Button – interaction	X	X	X	X	X	
	Button – state	X	X	X	X	X	
	Notification – status	X	X	X	X	X	
	Notification – type	X	X				
	Understanding feature functionalities				X	X	
	Raw NASA-TLX	42	21	84	23	6	35
	# of clicks (required: 4)	26	12	30	34	40	28 (14.2 %)

4.3 Detailed Results

Participant 1

The first participant did tasks 1 and 3 quickly and efficiently, without any observed difficulties as can be seen in Table 3. This was aligned with the subjective cognitive load scores. The number of clicks in these tasks was also the same as the required number of clicks. Thus, the participant was able to complete the tasks efficiently. Furthermore, regarding the effectiveness, the participant was able to fully achieve all task objectives. However, in tasks 2, 4 and 5 there were clear difficulties in the ease of use in various areas. The observed usability issues included all nine different themes which emerged during the analysis.

Table 3. Findings from the P1 usability evaluation session

	Raw NASA-TLX	Usability issues	Number of clicks	
			Test	Required
T1	18	-	1	1
T2	35	Distinguishing non-interactive elements Navigation – findability Navigation – interaction	12	2
T3	18	-	3	3
T4	35	Prototype issue Understanding feature functionalities	15	3
T5	42	Button – interaction Button – state Notification – status Notification – type	26	4

In T2 there were significant difficulties in accomplishing the task caused by three different usability issues in finding the correct view, interacting with the navigation and understanding which areas of the user interface are interactive and which are not. Moreover, the main issue was related to understanding where they currently are in the system from the navigational point of view. These usability challenges had a connection with the

experienced cognitive effort. Furthermore, the number of clicks was a lot bigger compared to the required number of clicks. This indicates an efficiency issue and further indicates clear usability issues on the task.

Similarly, in T4 the subjective cognitive load was on the same level as in T2 and the participant had observable usability issues in two different themes. One of the issues was caused by having difficulties understanding the feature functionality. Furthermore, slight confusion was caused by the prototype itself, as it did not provide real-time feedback for all functionalities. The number of clicks during the task also aligned with the cognitive load score and the usability challenges.

The highest cognitive load was reported in the T5, which inherently was the most strenuous task of the test. The task included interpreting information presented in the user interface which caused clear challenges, and clear difficulties were observed in interpreting the notification status and types. Moreover, it became prominent that the task was very demanding for the participant, as they reported a language barrier saying that they do not speak or understand English well. This can be assumed to affect the perceived cognitive effort also during the previous tasks. The number of clicks was also the highest in T5, which further supports the higher cognitive load and observed usability issues.

When assessing the cognitive load scores in comparison to the observed issues regarding the ease of use, there was a clear connection. While the subjective cognitive load scores were not extremely high in any of the tasks, they were noticeably higher in tasks 2, 4 and 5 in which the participant also experienced clear usability issues. Furthermore, in these tasks the efficiency was also worse compared to the other tasks where usability issues could not be observed.

In contrast, the participant reported lower scores in the experienced effort in T1 and T3, in which there were no usability issues observed. Nonetheless, some level of effort was reported even when no usability issues were found. There was observable uncertainty in multiple tasks during which the participant hovered their finger over the interface without doing any actions, seemingly unsure on how to proceed. Nevertheless, the participant successfully accomplished the objectives of all the tasks.

Participant 2

The second participant was able to accomplish 4 out of 5 task objectives, of which the T3 failed. Moreover, they reported higher cognitive load scores in the first three tasks than in the last two tasks, as seen in Table 4. These also aligned with the observed challenges in the system's usability. However, the number of different usability issues

did not seem to have a connection with the experienced cognitive load. This would suggest that in this case, P2 experienced higher required effort when having difficulties with the navigation of the interface than other types of usability issues. Moreover, the participant reported that their English language skills are not very good, and that it affected the understandability of the system. This can be assumed to influence the subjective cognitive load during the whole test session.

Table 4. Findings from the P2 usability evaluation session

	Raw NASA-TLX	Usability issues	Number of clicks	
			Test	Required
T1	42	Navigation – findability	25	1
T2	42	Button – state Navigation – findability	4	2
T3	42	Navigation – findability Understanding feature functionalities	9	3
T4	27	-	3	3
T5	21	Button – interaction Button – state Notification – status Notification – type	12	4

In T1 the participant seemed to be very lost from the beginning. When they did not find the desired information quickly, the participant reported that they would go through all different views of the interface just in case, as they attempted to find the desired information. This comment supports the observation of them being lost and having difficulties in navigating the interface. Furthermore, this corresponds with the cognitive load score being the highest that the participant gave during the evaluation session. Moreover, the efficiency was the worst of all the tasks.

In T2 the experienced cognitive load stayed at the same level as in the previous task. While the number of clicks was not extremely high compared to the required clicks, there were still a bit of issues with findability. On top of having issues with finding the correct information, the user had difficulties in understanding the button states during the task. As this task considers starting the crushing process, understanding the on and off

buttons on the interface is crucial. Therefore, it is expected that the cognitive load is high, even though the efficiency of completing the task seemed to be quite good.

The participants did not fully complete T3, although they were under the assumption that they had accomplished the task's objective. However, they noticed their error later during the test session and pointed out what they had missed in this task. The participant claimed that if it was a real-life situation they would have noticed their error, as the real machine would have given feedback. Had the prototype given real-time feedback to the participant, they most likely would have fully succeeded in T3. The efficiency, the observed usability challenges, and the experienced cognitive effort had a clear connection on this task.

While T5 is inherently more strenuous than the previous four tasks, the cognitive load score was the lowest in this test session. Furthermore, even though in T5 there were the biggest number of different usability issues observed, it was experienced as the most effortless task compared to others. However, the efficiency in T5 was not optimal, indicating that the efficiency had a relation with the observed usability findings.

When examining the number of clicks during the tasks, there seems to be quite a clear connection between efficiency, cognitive load and observed usability issues. For example, in T4 the number of clicks was the same as the required number of clicks, where also no usability issues were identified. However, the experienced cognitive load was the second lowest, even though no issues were found during the task. Nevertheless, the difference between the cognitive load in T4 and the lowest score which was given in T5 seems to be quite marginal, indicating that no definitive deductions can be made whether there were some underlying issues that would have increased the amount of experienced cognitive load during T4. Moreover, in the other four tasks where usability issues were found, the number of clicks was also a lot bigger than the required number of clicks. Thus, efficiency was worse when the usability challenges were present.

Participant 3

The third participant reported the highest cognitive load scores throughout the test session compared to the other participants as seen in Table 5. This aligned well with the observed attitudes, as the participant was most visibly frustrated and annoyed with the system throughout the test session. On top of this, they experienced a lot of difficulties in using the interface altogether. Moreover, the participant was the least effective, being able to achieve only 3 out of 5 tasks, from which they did not fully finish tasks 1 and 3.

In T1 the participant seemed very frustrated from the start. They seemed extremely dissatisfied with the mere idea that they would need to start navigating in the system to find information. Furthermore, when they did not find the correct information easily, the inefficiency seemed to cause a lot of negative feelings. The experienced effort aligns well with the observed usability issues and the efficiency of using the system.

In T2 the participant did not know whether they managed to complete the task or not. They also reported that they do not understand what is happening on the interface, which is extremely crucial finding as the task considered starting the crushing process. Furthermore, the difficulties could also be seen by observing their behavior. The participant attempted to do the same actions multiple times, indicating that the interactions and the state of the button were confusing. This also had an obvious connection with the cognitive load score, as it was the highest of all in T2. Furthermore, the number of clicks is higher than what is required to complete the task.

Table 5. Findings from the P3 usability evaluation session

	Raw NASA-TLX	Usability issues	Number of clicks	
			Test	Required
T1	81	Button – interaction Navigation – findability	16	1
T2	93	Button – interaction Button – state	6	2
T3	90	Understanding feature functionalities	7	3
T4	61	-	3	3
T5	84	Button – interaction Button – state Notification – status	30	4

Similarly to the previous task, the participant reported difficulties in understanding whether they did the right actions in T3. They also experienced the second highest effort during this task and did not fully achieve the objectives of the task. Furthermore, the number of clicks indicates that there were issues regarding efficiency.

While there were no observable usability issues in T4, the participant reported that they experienced relatively high cognitive effort. However, compared to their previous answers there is a clear decrease in the cognitive load, so in that perspective the results are clearly connected. Furthermore, as the participant had felt that using the system requires a lot of effort from them in previous tasks, it may affect their ratings even though they could use the system easily. This suggests that the overall experience may affect the experienced effort, even though they would have succeeded in the task perfectly and efficiently. Furthermore, this emphasizes the importance of satisfaction, which is one of the main aspects regarding usability.

The participant encountered multiple observable usability issues in T5, which can also be seen to be aligned with the experienced cognitive effort being again at the higher end of the scale. However, no clear connection can be found with the number of issues, as they reportedly experienced higher cognitive load in T2 where only two issues were present compared to three different issues in T5. Again, they reported that they do not understand what is presented on the interface while still being able to accomplish the task's objectives correctly. However, the number of clicks indicates that there are serious issues with using the interface efficiently in T5.

Overall, in this test session the participant's experienced cognitive load reflected quite well with the usability issues, as the reported amount of experienced effort was higher in the tasks where usability issues were present as well as in tasks which the participant was not able to accomplish. Moreover, the number of clicks was higher in those tasks. However, it is apparent that the amount of usability issues was not affecting the perceived cognitive load. Furthermore, in this case the participant was seemingly frustrated during every task, so the cognitive load scores may be affected by their overall mental state during the test.

Participant 4

The fourth participant experienced observable usability issues in most of the tasks, excluding T3. Moreover, the participant did not manage to accomplish the T2 objectives. However, they did not reportedly experience high cognitive load in all tasks where usability issues could be observed, or where the efficiency of use was suffering. As seen in Table 6, the scale of the experienced cognitive effort varied a lot.

There were distinct, observable usability issues in T1, which also reflects the number of clicks during the task. Even though multiple issues were observed in T1, and the participant was extremely lost when navigating in the system, their experienced cognitive load

was not rated high. Thus, there was no clear connection between the usability issues and in the cognitive load scores.

Table 6. Findings from the P4 usability evaluation session

	Raw NASA-TLX	Usability issues	Number of clicks	
			Test	Required
T1	28	Navigation – findability Navigation – interaction	33	1
T2	83	Button – interaction Button – state Distinguishing non-interactive elements	62	2
T3	8	-	4	3
T4	8	Button – interaction Prototype issue Understanding feature functionalities	16	3
T5	23	Button – interaction Button – state Notification – status Understanding feature functionalities	34	4

As established before, the participant did not manage to accomplish the objectives in T2, which I also reflected in the amount of usability issues observed during the task. While the participant did find the correct place quickly, they did not understand how specific elements on the interface work, they started to explore the whole interface in hopes of finding another place where that objective could be executed. This is also reflected in the number of clicks being the highest in T2 during this test session. The subjective cognitive load was also the highest in T2, and the task caused observable frustration. Therefore, there is a clear relationship between the experienced cognitive load, usability issues and efficiency.

One possible cause for the experienced low cognitive effort in T1 may be explained by the outcome of the task. Even though the participant experienced a lot of challenges in

navigating the system, they were able to accomplish the task's objective. This might explain why they experienced lower effort when retrospectively evaluating the required cognitive effort. In contrast to T2, which they could not accomplish, the subjective cognitive load score was extremely high. Furthermore, it is noteworthy that in T2 the participant managed to find the correct place for doing the desired actions, but was confused because they did not understand what information was presented in the interface and how the interface's functionalities should work. This caused them to start exploring different parts of the interface, as they started to question themselves and whether they had even found the right place to do the desired action.

In T4 the participant experienced some issues as the prototype's feedback was not working exactly as it would work with the real machine. However, they understood how it should work in real life setting. Due to the prototype issues, the participant started to explore other ways to reach the objective, even though the first approach was right. However, this revealed problems in understanding the interface's functionalities. Even though the participant seemed to struggle with the use of the interface, they reported the same amount of cognitive effort as in T3 where no usability issues were observed. Thus, the perceived cognitive score does not align well with the usability findings or with the number of clicks in this task.

Even though in T5 there were observable issues in usability, and the participant seemed to be quite confused during the task, the cognitive load score was only the second highest compared to other tasks. For example, in T1 the number of issues was same as in T5, but in T5 the subjective cognitive load was a bit lower. This may suggest that the type of usability issue might impact on how much effort one may experience during a task. For example, in T1 the biggest issue was related to navigating and finding the correct place within the system. In T5 that was not an issue, so in this case it could be assumed that if the participant is not able to find the correct place easily, it amounts to a higher cognitive load and frustrates the user a lot more compared to having difficulties in understanding what information is presented on the screen. However, the difference between the required effort is quite small, so nothing definitive can be deducted.

Participant 5

The final participant reported extremely distinct experiences of the required effort compared to the previous participants. Throughout the test, their experienced effort was very low according to the questionnaire results, as can be seen in Table 7. However, these results did not always align with the observed usability findings. Overall, there were

various issues observed in usability and efficiency, excluding T1 which the participant managed to do quickly and efficiently without experiencing excessive cognitive load.

Table 7. Findings from the P5 usability evaluation session

	Raw NASA-TLX	Usability issues	Number of clicks	
			Test	Required
T1	7	-	1	1
T2	3	Button – interaction	3	2
T3	7	Button – state	4	3
T4	9	Button – interaction Prototype issue Understanding feature functionalities	25	3
T5	6	Button – interaction Button – state Notification – status Understanding feature functionalities	40	4

In T2 there were observable challenges in understanding the interactions of button elements. Nevertheless, in this task, the participant reported that they experienced the lowest cognitive effort compared to any other task. While the usability issues were not causing massive problems during the task and they were able to complete the task objective, erroneous actions could be observed. The number of clicks was also close to the required number, indicating that there were no issues regarding efficiency as the difference to the required number of clicks was only marginal.

Similarly to the previous task, in T3 there were observable challenges with understanding button's states. However, the task was completed despite the observed unclarity and erroneous actions. Furthermore, it is notable that the participant experienced the same amount of cognitive effort as in T1, even though T3 seemed to cause more struggles while observing the use of the system.

In T4, it is notable that the participant achieved the task objective correctly at first, with minimum number of clicks. However, they did not understand that they had in fact accomplished the aim of the task and started to do various actions and figure out other possible solutions. The main reason for the confusion was caused by the lack of real-time feedback from the prototype. However, after they started to explore other ways to accomplish the task objective, there were clear issues in understanding various features on the interface. While the experienced cognitive load is not rated as high, it is still the highest rating from this test session. Nevertheless, when compared to the participant's other tasks, the difference is still marginal and may not indicate any definitive outcomes.

The final task also did not seem to require much cognitive effort from the participant according to the questionnaire results. However, the highest number of different usability issues were observed during the task, which can also be seen to reflect in efficiency. The participant had difficulties in understanding how to interact with various elements on the interface. When they tried to interact with them in a way that does not work, it increased the number of clicks. Furthermore, they tried erroneous interactions multiple times before changing the approach of interacting with the elements. However, the cognitive load score was lower than in T1 where no usability issues were observed. Therefore, when comparing the experienced effort to the usability issues or efficiency, no clear connection was found.

One reason for the low cognitive load scores can be explained by the overall attitude. In this test, the participant seemed to be very laid back and indifferent towards the system under evaluation. For example, in T4 they were not sure if they even did the task correctly, but it did not seem to bother them. They wanted to move on to the next one while appearing careless of the outcome. This can account for the ratings given in the questionnaire. The test setting might have also influenced the participant's attitude, as the test was conducted with a prototype, and the lack of the real machine can lower the amount of experienced effort.

5. DISCUSSION

This chapter discusses the findings, limitations and future research areas, and the strengths and contributions of the study. First, the main findings of the research are introduced and discussed, focusing on what can be interpreted from the results. Moreover, the findings are discussed in combination of the selected method for measuring the cognitive load. After that, the identified limitations of this study are acknowledged, and possible further research areas are addressed. Finally, the strengths of the research and its contributions are discussed at a general level.

5.1 Research Findings

The aim of this research was to study the experienced cognitive load and observe how much effort the users need to accompany while using the product, in relation to possible usability challenges that can be observed. Therefore, to collect the desired data, the study was carried out by conducting task-oriented usability tests with a high-fidelity prototype of a new heavy machinery interface, focusing on the main operational functionalities. The participants' cognitive load was measured by utilizing a Finnish translation of the NASA-TLX questionnaire, which was used for assessing the experienced cognitive effort subjectively. As a result, multiple connections could be identified between usability and the experienced cognitive load, while also revealing areas where clear relationship could not be found.

One of the main research findings indicates that there is a connection between the experienced cognitive load and the different types of usability issues. When specific types of usability challenges emerged throughout the usability evaluation, the cognitive load was often rated as higher among the participants. For example, if a participant was unable to navigate the system efficiently, it resulted in observable frustration, which in turn contributed to an increased level of experienced cognitive load. These results further support previous studies where a similar connection between navigational issues and higher levels of cognitive load have been recognized (Shi et al., 2021). Furthermore, the findings may indicate that navigational problems are occupying more memory chunks from the working memory in the information processing model. However, in this research, the results also indicated that higher cognitive load could be identified when usability issues regarding the visibility of system status were present. When the participants had difficulties in understanding whether the machine's devices were on or off due to

confusion caused by button statuses, the experienced cognitive load seemed to increase in parallel.

Furthermore, the level of efficiency seemed to have a relation with cognitive load in most of the tasks. When the observed number of clicks during a task was remarkably higher than the required number of clicks for task completion, indicating poor efficiency, the required cognitive load seemed to increase. In contrast, the participants reported less experienced effort in most of the tasks which they managed to accomplish efficiently, which strengthens the consistency with previous research (Al-Saud, 2023; Yurko et al., 2010). Nevertheless, in a few tasks the connection between efficiency and cognitive load was not as distinct, and despite the participants being able to accomplish the tasks relatively efficiently, higher cognitive load scores were reported.

When considering the different dimensions of usability, the importance of satisfaction became prominent during the research. As said, when the participants encountered usability issues, they often seemed to experience higher levels of frustration in addition. A few occurrences in the usability evaluation sessions indicated that if a participant was visibly annoyed by the system due to previous challenges in the ease of use, they reported higher cognitive load in the next task as well even when there would not be any observable usability issues present, and they accomplished the task efficiently. Kliegel et al. (2003) concluded in their study that existing negative emotions may increase the inherent cognitive load in information processing, which allocation would otherwise be reserved and consumed by doing an activity. Therefore, if the user is stressed, or in another way in a negative mental state before even starting to use the system, their experienced cognitive load may inherently be higher. Thus, the higher cognitive load in the study results might be caused by the previous, negative experiences of using the system. Furthermore, the lack of association between cognitive load and usability challenges in some tasks may be explained by the level of satisfaction. Additionally, while in this study the connection between cognitive load and emotional state became prevalent by studying the effort through subjective measurement, the same indications have been studied with physiological measures examining electrodermal activity, resulting in a clear correlation between mental state and cognitive effort (Nourbakhsh, 2017).

Most importantly, the results imply that assessing only the amount of experienced effort during usability evaluation with a subjective measure such as the NASA-TLX questionnaire does not reveal all potential areas where challenges in usability could be encountered. For example, in some tasks the experienced cognitive load was relatively high, even though no usability issues could be observed. Similarly, lower cognitive load scores do not necessarily indicate that using the product under evaluation would not have any

difficulties or challenges. Multiple studies have concluded similar findings, indicating that measuring the mental workload in conjunction to usability does not examine the same phenomenon, but in fact studies a somewhat different viewpoint of the same concept (Matthews et al., 2020; McKendrick & Cherry, 2018; Longo, 2018). The findings in this research further align with previous research, concluding that a participant may as well experience low cognitive effort even when the efficiency is not optimal. Furthermore, as the lack of clear connection between the number of different usability issues identified in a task and the experienced cognitive effort became prevalent in this study, the results suggest that the underlying issues in the experienced cognitive load may not be distinctly related to the findings from usability evaluation. Moreover, the lack of connection further emphasizes how the type of usability issue affects the effort. In addition, while some outliers in the data were revealed during the analysis, in most cases the aforementioned connections could be identified. However, when the cognitive load assessment is combined with other metrics, such as the number of clicks and qualitative insights about the usability issue types encountered, measuring the subjective cognitive load brings additional, meaningful viewpoints to usability evaluation especially in terms of satisfaction.

It is notable that there may be variation in the self-evaluation of the experienced cognitive load. As concluded before, self-assessment methods may not always reflect the actual performance as people might not be able to do realistic estimations (Karpen, 2018). Instead, the subjective evaluation may be biased into both more optimistic and pessimistic outlooks depending on the individual (Deffuant et al., 2024; Gadsby & Hohwy, 2023). This could also be seen in the results of this study. Some participants reported their cognitive load consistently at the higher end of the scales even when no usability issues could be detected, while other participants were more moderate in their assessments throughout the test. However, when comparing the individual participants' results to their overall answers during the test, the increase or decrease in the cognitive load could be often identified.

While the cognitive load and usability seemed to be more related constructs than being straightforwardly connected to each other, the extraneous cognitive load caused by poor design choices could be identified. Moreover, it could expose the user to erroneous actions from the beginning, which further emphasizes the importance of usability. While this study further revealed the importance of satisfaction, it is noteworthy that the interface was not even used in the real context of use, which may potentially increase the required effort and strain in the use of the human-machine interface. Moreover, there was an observable connection between high cognitive load and satisfaction. Thus, the

study further suggests that satisfaction as a part of the concept of usability should be considered equally important as efficiency and effectiveness.

5.2 Limitations and Future Research

There were a few limitations that could be identified in this study. First, it should be acknowledged that the sample size of five participants is limited. The nature of the potential participants' work can be quite fast paced, and the changes in their work location and availability during shifts can change very quickly. This limited the opportunities to participate and commit to the whole test session. Furthermore, conducting the evaluation itself on the customer's premises consumes the participant's time and resources, which further restricted the possibility of recruiting suitable candidates. However, in qualitative usability testing five test sessions are generally considered as a sufficient amount to find out the most critical usability issues as the study by Nielsen & Landauer (1993) concluded. Nevertheless, it should be noted that the number of participants impacts the cognitive load studies as it does not represent a wide user group. Therefore, the relatively small sample size considering cognitive load studies should be acknowledged when interpreting the results, as it can convey a slightly narrow view of the perceived overall cognitive load. Moreover, as the overall self-assessment of cognitive load showed fluctuation, the number of participants limits the possibilities to analyze more deeply the reasons behind the variation. This also poses a further research topic to enhance the reliability of the study by verifying the results with a larger sample size.

Another limitation of the study considers the controlled environment and the lack of real context of use. The test was conducted with a high-fidelity Figma prototype of the human-machine interface while being indoors and not in the real use environment. While the prototype provided close resemblance to the actual display in size and functionality, the real context of use may be vastly different when using the product in actual work scenario. For example, the environmental factors, such as different weather conditions, were not affecting this test session. Thus, the test may not fully reflect the amount of required cognitive effort when using the interface of the actual machine as the possible external strain from the real context of use could not be accounted for in the test. In a study conducted by Krebs (2024) it was concluded that an increase of even 1 °C in the temperature was already affecting the users' cognitive performance above the threshold of 16.5 °C, leading to lower efficiency and higher risk of errors. This can be assumed to be true also in the use of the interface that was tested in this study, further emphasizing the importance of testing the product in real context. Moreover, as the evaluation was conducted with a prototype, no harm could be done during the test sessions and erroneous

actions would not have any consequences. This may impact the level of experienced cognitive load, as the external strain was not present. However, this would not be the case when using the real machine. Therefore, to enhance the validity of measuring the cognitive load during the use of the human-machine interface, further study is required in the real context of use.

The NASA-TLX questionnaire examines the cognitive load by assessing the participant's experienced effort from six different viewpoints, thus having six different subscales in the questionnaire. However, this study did not examine the subscales individually but relied on the average scores calculated from the questionnaire. One aspect for further studies could consider analyzing how the different subscales might have been impacted, and if any results from distinct subscales would suggest a stronger connection between cognitive load and usability evaluation findings than other subscales. Furthermore, considering the overall user experience, the satisfaction dimension of usability could be inspected through another measurement, such as user experience questionnaires or usability questionnaires. This way clearer insights on the overall satisfaction could be gathered, and the connection between satisfaction and cognitive load could be examined in the context of usability. Moreover, to increase the validity of the cognitive load results, additional measures could be introduced in future research. The self-assessment method showed a bit of variation in the scaling, therefore the cognitive load could be studied also by utilizing different physiological measures, such as eye-tracking or electrodermal activity, to examine whether the deviations in the results could be possibly mitigated.

5.3 Strengths and Contributions

One of the primary contributions of this research is examining the connection between cognitive load and usability in the context of complex human-machine interface. Moreover, a mixed-method approach was utilized for the purpose of triangulating the data to enhance the credibility of the results. Furthermore, while the focus of the case study was related to human-machine interfaces in the field of heavy machinery, the learnings from this research can be applied to other areas as well, which further demonstrate the contributions of the research.

From the theoretical point of view, the study advances the understanding of the relationship between usability and cognitive load in intricate systems, and how measuring cognitive effort may bring valuable insights into the means of how information processing flow can be impacted while using an inherently complex interface. Especially in the field of work where the system itself may cause strain, and the intrinsic cognitive load might already be high, the additional complexity caused by the interface itself affects how well

users can interact with the system in the first place. Therefore, the importance of understanding the extraneous cognitive load is emphasized in this study. Furthermore, the research findings support the existing views on how usability is not only connected to the ability to complete tasks but is a multifaceted concept where the cognitive strain should be noted as an impactful concept related to the ease of use, while still being a separate construct (Longo, 2018).

Inspecting the research from the methodological point of view, it revealed how these two different methods work together in usability evaluation. While these methods have been utilized together before, the study further indicated the applicability of using the methods in the prototype testing with human-machine interfaces. Moreover, the variability and depth of the results brought out important aspects regarding the connection between usability and cognitive load, indicating also faults in the self-assessment and pointing out the importance of the mental state during the evaluation.

Finally, the importance of evaluating the cognitive load during usability evaluation became prominent in this study. In some cases, even though usability issues could not be identified during a task, there were clear indications of high cognitive load and frustration. It is known that high cognitive load exposes users to hazardous situations by increasing the possibility of them performing erroneous actions. This study further revealed how it is necessary to understand what aspects of user interfaces may cause additional cognitive strain from the usability point of view, so that these situations can be mitigated to ensure safety and user satisfaction. Furthermore, these findings may apply to other fields as well, and not only to human-machine interfaces.

6. CONCLUSIONS

The aim of this research was to examine the connection between cognitive load in the context of usability. A mixed-method study was conducted to find out how subjective cognitive load aligns with the observed usability evaluation results, and how measuring the intrinsic cognitive load can indicate the presence of usability issues. The research questions were the following:

1. How does cognitive load align with observed usability challenges during usability evaluation of a human-machine interface?
2. How well can measuring cognitive load indicate difficulties in usability?

The results revealed that there is no obvious connection between the amount of experienced cognitive load and observed usability challenges during the use of a human-machine interface. The amount of reported cognitive load did not in every case clearly increase when usability issues were encountered, and conversely in some cases cognitive load was reported relatively high even when no issues could be observed. Moreover, the number of usability issues did not seem to have a clear effect on the level of experienced cognitive effort.

Regarding the efficiency of use and cognitive load, a connection could be found on some level, which is aligned with previous research on the topic. In most cases, the cognitive load was reported to be higher when the number of clicks in relation to required number of clicks increased significantly. However, deviation could be found, so future studies should be conducted to enhance the validity of the results.

Moreover, the findings suggest that the type of encountered usability challenges may affect the experienced cognitive load. When specific types of usability issues were encountered during a task the cognitive load was often higher, and conversely the cognitive load was in some cases lower even when multiple usability challenges could be observed. However, as the sample size of the study was limited, it should be further studied to validate the findings.

The research revealed a connection between the mental state and cognitive load. When observable negative feelings were present, it seemed to increase the experienced cognitive load. Similarly, more optimistic attitudes were often connected with lower levels of experienced effort. Thus, the results further indicate fault in the self-assessment of

cognitive load, which should be further studied to find out whether no effort was experienced or whether the optimistic outlook affected the results.

In conclusion, while no clear and definitive connection could be found between the subjective cognitive load and usability, the results suggest that these concepts are nonetheless related to each other to some degree, which aligns well with previous research. Therefore, measuring high levels of subjective cognitive load during usability evaluation does not straightforwardly indicate issues in usability. Conversely, the lack of cognitive load does not mean that the user would not encounter any usability issues.

REFERENCES

- Al-Saud, L. M. (2023). Simulated skill complexity and perceived cognitive load during preclinical dental training. *European Journal of Dental Education*, 27(4), 992–1003. <https://doi.org/10.1111/eje.12891>
- Atkinson, R. C., & Shiffrin, R. M. (2024). Reprint of: Human memory: A proposed system and its control processes. *Journal of Memory and Language*, 136, 104479-. <https://doi.org/10.1016/j.jml.2023.104479>. (Original work published 1968)
- Baddeley, A. (1997). *Human memory: theory and practice*. Psychology Press.
- Baddeley, A. D. (2007). *Working memory, thought, and action*. Oxford University Press.
- Baddeley, A., Taylor, S., Fiske, S., & Schacter, D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, 63(1), 1–29. <https://doi.org/10.1146/annurev-psych-120710-100422>
- Baddeley, A. D., Eysenck, M. W., & Anderson, M. C. (Michael C.). (2020). *Memory* (Third edition.). Routledge.
- Barshi, I., Degani, A., Mauro, R., & Mumaw, R. J. (2024). Models of Human-Automation Systems: Initial Analysis of the Boeing 737MAX Design. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 68(1), 835–840. <https://doi.org/10.1177/10711813241279805>
- Clarke, M. A., Schuetzler, R. M., Windle, J. R., Pachunka, E., & Fruhling, A. (2020). Usability and cognitive load in the design of a personal health record. *Health Policy and Technology*, 9(2), 218–224. <https://doi.org/10.1016/j.hlpt.2019.10.002>
- Collet, C., Salvia, E., & Petit-Boulanger, C. (2014). Measuring workload with electrodermal activity during common braking actions. *Ergonomics*, 57(6), 886–896. <https://doi.org/10.1080/00140139.2014.899627>
- Cowan, N. (1998). Visual and auditory working memory capacity. *Trends in Cognitive Sciences*, 2(3), 77–77. [https://doi.org/10.1016/S1364-6613\(98\)01144-9](https://doi.org/10.1016/S1364-6613(98)01144-9)
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, 24(1), 87–114. <https://doi.org/10.1017/S0140525X01003922>
- Deffuant, G., Roubin, T., Nugier, A., & Guimond, S. (2024). A newly detected bias in self-evaluation. *PloS One*, 19(2), e0296383–e0296383. <https://doi.org/10.1371/journal.pone.0296383>
- DiDomenico, A., & Nussbaum, M. A. (2008). Interactive effects of physical and mental workload on subjective workload assessment. *International Journal of Industrial Ergonomics*, 38(11), 977–983. <https://doi.org/10.1016/j.ergon.2008.01.012>
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working Memory, Short-Term Memory, and General Fluid Intelligence: A Latent-Variable Approach. *Journal of Experimental Psychology. General*, 128(3), 309–331. <https://doi.org/10.1037/0096-3445.128.3.309>

- Frazier, S., Pitts, B. J., & McComb, S. (2022). Measuring cognitive workload in automated knowledge work environments: a systematic literature review. *Cognition, Technology & Work*, 24(4), 557–587. <https://doi.org/10.1007/s10111-022-00708-0>
- Gadsby, S., & Hohwy, J. (2023). Incentivising accuracy reduces bias in the imposter phenomenon. *Current Psychology (New Brunswick, N.J.)*, 42(32), 27865–27873. <https://doi.org/10.1007/s12144-022-03878-2>
- Gao, Q., Wang, Y., Song, F., Li, Z., & Dong, X. (2013). Mental workload measurement for emergency operating procedures in digital nuclear power plants. *Ergonomics*, 56(7), 1070–1085. <https://doi.org/10.1080/00140139.2013.790483>
- Goel, P., Datta, A., & Mannan, M. S. (2017). Industrial alarm systems: Challenges and opportunities. *Journal of Loss Prevention in the Process Industries*, 50, 23–36. <https://doi.org/10.1016/j.jlp.2017.09.001>
- Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.) *Human Mental Workload*. Amsterdam: North Holland Press.
- Hautamäki, E., Kinnunen, U.-M., & Palojoki, S. (2017). Health information systems' usability-related use errors in patient safety incidents. *Finnish Journal of eHealth and eWelfare*, 9(1), 6-. <https://doi.org/10.23996/fjhw.60763>
- Harvey, C., Stanton, N. A., Pickering, C. A., McDonald, M., & Zheng, P. (2011). A usability evaluation toolkit for In-Vehicle Information Systems (IVISs). *Applied Ergonomics*, 42(4), 563–574. <https://doi.org/10.1016/j.apergo.2010.09.013>
- Hassenzahl, M., & Tractinsky, N. (2006). User experience - a research agenda. *Behaviour & Information Technology*, 25(2), 91–97. <https://doi.org/10.1080/01449290500330331>
- Hertzum, M., Hansen, K. D., & Andersen, H. H. K. (2009). Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology*, 28(2), 165–181. <https://doi.org/10.1080/01449290701773842>
- International Organization for Standardization. (2018). Ergonomics of human-system interaction. Part 11: Usability: Definitions and concepts (ISO Standard No. 9241-11:2018). <https://www.iso.org/standard/63500.html>
- International Organization for Standardization. (2019). Ergonomics of human-system interaction. Part 210: Human-centred design for interactive systems (ISO Standard No. 9241-210:2019). <https://www.iso.org/standard/77520.html>
- International Society of Automation. (2015). Human Machine Interfaces for Process Automation Systems (ANSI/ISA-101.01-2015). International Society of Automation.
- Jaiswal, D., Chowdhury, A., Banerjee, T., & Chatterjee, D. (2019). Effect of Mental Workload on Breathing Pattern and Heart Rate for a Working Memory Task: A Pilot Study. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, 2202–2206. <https://doi.org/10.1109/EMBC.2019.8856458>

- Karpen, S. C. (2018). The Social Psychology of Biased Self-Assessment. *American Journal of Pharmaceutical Education*, 82(5), 441–448. <https://doi.org/10.5688/ajpe6299>
- Khawaja, M. A., Chen, F., & Marcus, N. (2014). Measuring Cognitive Load Using Linguistic Features: Implications for Usability Evaluation and Adaptive Interaction Design. *International Journal of Human-Computer Interaction*, 30(5), 343–368. <https://doi.org/10.1080/10447318.2013.860579>
- Kliegel, M., Horn, A. B., & Zimmer, H. (2003). Emotional after-effects on the P3 component of the event-related brain potential. *International Journal of Psychology*, 38(3), 129–137. <https://doi.org/10.1080/00207590344000006>
- Krebs, B. (2024). Temperature and Cognitive Performance: Evidence from Mental Arithmetic Training. *Environmental & Resource Economics*, 87(7), 2035–2065. <https://doi.org/10.1007/s10640-024-00881-y>
- Longo, L. (2018). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PloS One*, 13(8), e0199661–e0199661. <https://doi.org/10.1371/journal.pone.0199661>
- Mandler, G. (2002). Origins of the cognitive (r)evolution. *Journal of the History of the Behavioral Sciences*, 38(4), 339–353. <https://doi.org/10.1002/jhbs.10066>
- Matthews, G., De Winter, J., & Hancock, P. A. (2020). What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. *Theoretical Issues in Ergonomics Science*, 21(4), 369–396. <https://doi.org/10.1080/1463922X.2018.1547459>
- McKendrick, R. D., & Cherry, E. (2018). A Deeper Look at the NASA TLX and Where It Falls Short. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1), 44–48. <https://doi.org/10.1177/1541931218621010>
- Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97. <https://doi.org/10.1037/h0043158>
- Moore, E., Laiti, L., & Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. *Nature Neuroscience*, 6(2), 182–189. <https://doi.org/10.1038/nn996>
- Nachreiner, F., Nickel, P., & Meyer, I. (2006). Human factors in process control systems: The design of human–machine interfaces. *Safety Science*, 44(1), 5–26. <https://doi.org/10.1016/j.ssci.2005.09.003>
- Nielsen, J. (1993). *Usability engineering*. Boston: AP Professional.
- Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. *INTERCHI '93: Conference Proceedings: Bridges between Worlds*, 206–213. <https://doi.org/10.1145/169059.169166>
- Norman, D. (2013). *The Design of Everyday Things: Revised and Expanded Edition* (Rev. and expanded ed.). Basic Books.

- Nourbakhsh, N., Chen, F., Wang, Y., & Calvo, R. A. (2017). Detecting users' cognitive load by galvanic skin response with affective interference. *ACM Transactions on Interactive Intelligent Systems*, 7(3), 1–20. <https://doi.org/10.1145/2960413>
- Oliver, H., Kostkova, P., & de Quincey, E. (2010). Data Triangulation in a User Evaluation of the Sealife Semantic Web Browsers. *Electronic Healthcare*, 27, 80–87. https://doi.org/10.1007/978-3-642-11745-9_13
- Olivers, C. N. L. (2011). Long-term visual associations affect attentional guidance. *Acta Psychologica*, 137(2), 243–247.
- Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist*, 38(1), 63–71. https://doi.org/10.1207/S15326985EP3801_8
- Palviainen, J., Väänänen-Vainio-Mattila, K., & Kurosu, M. (2009). User Experience in Machinery Automation: From Concepts and Context to Design Implications. In *Human Centered Design* (Vol. 5619, pp. 1042–1051). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-02806-9_119
- Pan, T., Wang, H., Si, H., Li, Y., Li, G., & Zhu, Y. (2024). Research on identification of flight cadets' cognitive load based on multi-source physiological data and CGAN-DBN model. *Ergonomics*, 1–19. <https://doi.org/10.1080/00140139.2024.2380340>
- Paz, F., & Pow-Sang, J. A. (2014). Current Trends in Usability Evaluation Methods: A Systematic Review. *2014 7th International Conference on Advanced Software Engineering and Its Applications*, 11–15. <https://doi.org/10.1109/ASEA.2014.10>
- Plass, J. L., Moreno, R., & Brünken, R. (2010). *Cognitive load theory*. Cambridge University Press.
- Purves, D., Cabeza, R., Huettel, S. A., LaBar, K. S., Platt, M. L., & Woldorff, M. G. (2013). *Principles of cognitive neuroscience* (2nd ed.). Sinauer Associates.
- Ren, Z., Guo, F., Li, M., Lyu, W., & Duffy, V. G. (2024). The effect of in-vehicle agent embodiment on drivers' perceived usability and cognitive workload: Evidence from subjective reporting, ECG, and fNIRS. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 34(4), 325–337. <https://doi.org/10.1002/hfm.21030>
- Said, S., Gozdzik, M., Roche, T. R., Braun, J., Roessler, J., Kaserer, A., Spahn, D. R., Noethiger, C. B., & Tscholl, D. W. (2020). Validation of the raw national aeronautics and space administration task load index (NASA-TLX) questionnaire to assess perceived workload in patient monitoring tasks: Pooled analysis study using mixed models. *Journal of Medical Internet Research*, 22(9), e19472–e19472. <https://doi.org/10.2196/19472>
- Salvador, C., Nakasone, A., & Pow-Sang, J. A. (2014). A systematic review of usability techniques in agile methodologies. *Proceedings of the 7th Euro American Conference on Telematics and Information Systems*, 1–6. <https://doi.org/10.1145/2590651.2590668>
- Sharp, H., Rogers, Y., & Preece, J. (2023). *Interaction design: beyond human-computer interaction* (6th edition). Wiley.

- Shi, A., Huo, F., & Han, D. (2021). Role of Interface Design: A Comparison of Different Online Learning System Designs. *Frontiers in Psychology, 12*, 681756–681756. <https://doi.org/10.3389/fpsyg.2021.681756>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. [https://doi.org/10.1016/0364-0213\(88\)90023-7](https://doi.org/10.1016/0364-0213(88)90023-7)
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295–312. [https://doi.org/10.1016/0959-4752\(94\)90003-5](https://doi.org/10.1016/0959-4752(94)90003-5)
- Sweller, J. (2010). Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educational Psychology Review, 22*(2), 123–138. <https://doi.org/10.1007/s10648-010-9128-5>
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory* (1st ed. 2011.). Springer New York. <https://doi.org/10.1007/978-1-4419-8126-4>
- Tan, W., Liu, D., & Bishu, R. (2009). Web evaluation: Heuristic evaluation vs. user testing. *International Journal of Industrial Ergonomics, 39*(4), 621–627. <https://doi.org/10.1016/j.ergon.2008.02.012>
- Watrin, J. P., & Darwich, R. (2012). On Behaviorism in the Cognitive Revolution: Myth and Reactions. *Review of General Psychology, 16*(3), 269–282. <https://doi.org/10.1037/a0026766>
- Yurko, Y. Y., Scerbo, M. W., Prabhu, A. S., Acker, C. E., & Stefanidis, D. (2010). Higher mental workload is associated with poorer laparoscopic performance as measured by the NASA-TLX tool. *Simulation in Healthcare: Journal of the Society for Medical Simulation, 5*(5), 267–271. <https://doi.org/10.1097/SIH.0b013e3181e3f329>
- Zimmermann, J. F., Moscovitch, M., & Alain, C. (2017). Long-term memory biases auditory spatial attention. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 43*(10), 1602–1615. <https://doi.org/10.1037/xlm0000398>

APPENDIX A: TRANSLATED NASA-TLX QUESTIONNAIRE

The NASA Task Load Index (NASA-TLX)

Ole hyvä ja merkitse kuhunkin asteikkoon **X** omaa tuntemustasi parhaiten kuvaavaan viivan päälle. Arvioi vain omaa tuntemustasi äskeisen tehtävän aikana. Kiitos!

1. Ajattelun vaativuus

Kuinka paljon ajattelua ja hahmottamista tehtävä vaati (esimerkiksi päättelyä, etsimistä tai muistamista)?



2. Fyysinen vaativuus

Kuinka paljon fyysistä aktiivisuutta tehtävä vaati (esim. klikkailua)? Oliko tehtävä vaivaton tai työläs?



3. Ajallinen vaativuus

Kuinka paljon aikapainetta tunsit tehtävän aikana? Oliko tahti hidas ja verkkainen vai nopea ja hektinen?



4. Oma suoriutuminen

Kuinka onnistunut suorituksesi oli omasta mielestäsi? Onnistuitko tavoitteiden saavuttamisessa?



5. Vaivannäkö

Kuinka paljon vaivaa (henkistä ja fyysistä) tehtävän suorittaminen vaati?



6. Turhautuneisuus

Kuinka epävarma, lannistunut, ärtynyt tai stressaantunut olit tehtävän aikana?

