The Influence of Usability on Cognitive Load and Eye-movements

A Study of the Potential of Eye-tracking in Usability Evaluation

Anton Jigsved & Max Mauritsson

LUND
UNIVERSITY

The Influence of Usability on Cognitive Load and Eye-movements

A Study of the Potential of Eye-tracking in Usability Evaluation

Published by
Department of Design Sciences
Faculty of Engineering LTH, Lund University
P.O. Box 118, SE-221 00 Lund, Sweden

Subject: Interaction Design (MAMM01)

Supervisor: Johanna Persson
Co-supervisor: Marcus Nyström
Examiner: Christofer Rydenfält

Abstract

With its increased availability on the commercial market, *eye-tracking* has become a widely accessible system for analysing eye-movements. This has opened up new possibilities in evaluation of *usability* in *human-computer interaction*, which has had a history of being expensive, time-consuming and often performed based on poorly documented standards and objectives.

Previous research has indicated the ability of *eye-tracking* to evaluate a users' *cognitive load* based on eye-movements. This research suggests that there exists a possibility to utilise this technology in evaluation of *usability*.

In this master thesis, an experiment involving 30 participants was conducted to examine the influence of *usability* on *cognitive load* and eye-movements. This was done by letting the participants interact with three interface-prototypes. Two of which violated some *fundamental design principles*, whilst the remaining one was designed according to these principles and acted as a point of reference.

The hypothesis was that the flawed designs would result in interfaces of poor *usability*, contributing to a higher *cognitive load* during usage. This would in turn be objectively reflected in some specific metrics derived from the recorded *eye-tracking* data.

While no differences in perceived cognitive load was found, one of the flawed prototypes, which intentionally disregarded the principles of proximity and functional grouping, was distinguished from the two others based upon three specific *eye-tracking* metrics. Therefore, it is suggested that *eye-tracking* data has the potential to objectively reflect the *usability* of an interface. However, further research is required to investigate under what circumstances our results generalise to other designs.

Sammanfattning

Med sin ökade tillgänglighet på den kommersiella marknaden har *eye-tracking* blivit ett allmänt tillgängligt system för att analysera ögonrörelser. Detta har öppnat upp för nya möjligheter för utvärdering av *usability* inom *human-computer interaction*, som tidigare har varit dyrt, tidskrävande och ofta utfört baserat på dåligt dokumenterade standarder och mål.

Tidigare forskning har visat att *eye-tracking* har förmåga att utvärdera användarnas *kognitiva belastning* baserat på ögonrörelser. Denna forskning tyder på att det finns en möjlighet att använda detta system för utvärdering av *usability*.

I detta examensarbete genomfördes ett experiment med 30 deltagare för att undersöka hur *usability* påverkar *kognitiv belastning* och ögonrörelser. Detta gjordes genom att låta deltagarna interagera med tre gränssnittsprototyper. Två av dessa prototyper hade avsiktliga brister baserat på *fundamentala designprinciper* medan den återstående följde dessa principer och användes som referens.

Hypotesen var att de bristfälliga designerna skulle resultera i gränssnitt med dålig *usability*, vilket skulle bidra till en ökad *kognitiv belastning* under användning. Detta skulle i sin tur reflekteras objektivt i vissa specifika mätvärden härledda från inspelad data från *eye-tracking*.

Trots att inga skillnader i upplevd kognitiv belastning identifierades kunde den bristfälliga prototypen, som avsiktligt bortsåg från principerna om närhet och funktionell gruppering, särskiljas från de två andra baserat på tre specifika mätvärden. Därför föreslås det att *eye-tracking* har potentialen att objektivt reflektera *usability* hos ett gränssnitt. Dock krävs mer forskning för att undersöka under vilka förutsättningar dessa resultat är giltiga.

Preface

This master thesis was carried out at Lund University between January and June of 2021. It was performed at the department of Design Sciences at the Faculty of Engineering, LTH. The study was performed in order to further investigate the possibilities of using eye-tracking to evaluate usability based on cognitive load.

The division of labour in this master thesis was equal between the two authors. Initially, Anton Jigsved focused on the design aspect of the project while Max Mauritsson focused on the technical aspects and planning of the experiment. After the experimental phase, the authors cooperated in analysis of the data and the writing of the report.

The authors would first and foremost like to direct their gratitude towards their two supervisors Johanna Persson and Marcus Nyström for their support, feedback, positive attitude and engagement throughout the entire process. The authors also gracefully acknowledge Lund University Humanities Lab and the Department of Design Sciences for providing the equipment needed to perform the experiment. Additionally, the authors would like to thank all participants for sacrificing some of their precious time to contribute to science and the graduation of two happy fools. Gratitude is also extended to the staff at IKDC for their patience and showing the importance of pragmatics. Finally, the authors acknowledge the importance of role models such as Andrea Pirlo in displaying elegance and grace both on and off the pitch.

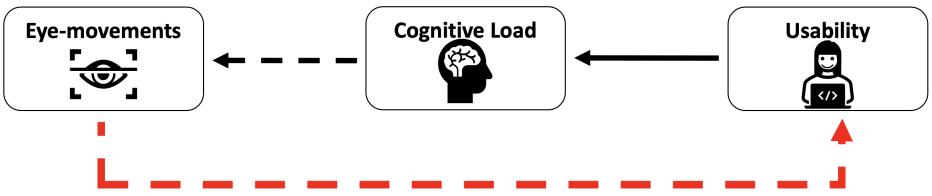
Lund, June 2021
Anton Jigsved & Max Mauritsson

Popular Science Summary

Eye-tracking - the Future of Usability Evaluation?

The movements of the eyes can tell a lot about our subconscious processing. With eye-tracking, interaction-designers have a powerful tool with which to unlock the secrets of the human mind and its cognitive processes. These processes are fundamental to the function of any human-computer interface and vital for overall usability. The results from this study show that eye-tracking can be used to distinguish between interfaces of varying design quality. Eye-tracking therefore has the potential to revolutionise the field of interface-design.

A usable product should guide the user during interaction and be designed in a way that hinders the user from committing mistakes. With the ongoing digitisation of all aspects of society, the importance of good interface-design can not be overstated. However, measuring usability is a complicated, time-consuming and often expensive process focused around subjective experiences. Since subjective experiences rarely tells the whole truth, this field is in need of a renaissance. Eye-tracking has the potential to lead this renaissance. By tracking the movements of the eyes, this technology supplies valuable information about how a interface is perceived and interpreted. This allows for instantaneous and objective analysis of the interaction that can offer easy to perform usability evaluation.



In this study, eye-tracking was used during 30 participants' interaction with three interface-prototypes of varying usability. Here, two of the prototypes violated established design principles whilst the remaining one acted as a point of reference. Using metrics derived from the eye-tracking data one of the flawed interface-prototypes was distinguished from the reference. This finding is especially interesting since an subjective assessment of the participants' experience showed no perceived difference in interface-quality between the flawed interface and the reference. This reinforces the capability of eye-tracking in capturing subconscious processes.

This study has only scratched the surface of the potential of eye-tracking. The findings encourage further, more comprehensive, research within this field. The method used also show that eye-tracking can be implemented and adjusted to fit any environment seamlessly. This offers a huge opportunity for a wide variety of real-life applications. If the eyes are truly the mirrors of the soul, eye-tracking could lead a renaissance in interface-design with the potential to revolutionise the entire field of usability evaluation.

Table of Contents

Chapter 1
Introduction

With the ongoing digitisation of nearly all aspects of modern societies, it is fair to say that we are living in, and have been for quite some time, a digital revolution. While this surely have facilitated many aspects of human life, there are no shortage of real-life examples where the function and design of digital solutions have contributed to stress and confusion. If such effects occur in the wrong environments, such as healthcare or military, they could lead to dire consequences.

One such example is the situation that occurred on the 13th of January 2018 on Hawaii. On this day, an employee of the Hawaii Emergency Management Alert was tasked to initiate a internal test of the emergency missile warning system. A system that in case of an actual threat can be utilised to send alerts to the public. The internal test that was to be initiated checks the function of the system without actually sending an alert to the public. However, due to inadequate design of the interface, the employee instead pressed the wrong option, sending out an actual alert-message warning people to immediately seek shelter [1].

This system has since then become a famous example of the real-life consequences of lacking design and poor usability. A usable product, i.e. a product with good usability, should give the user sufficient cues on how to use the system and be robust enough to avoid mistakes such as these from occurring [2]. The usability of a product is highly dependent on the design of the product. A helpful tool when designing a usable product is to utilise established design principles. These principles combine knowledge from several fields within design to generate concepts fundamental to the design of a usable product. One aspect addressed by design principles in order to achieve a good usability is minimising the cognitive load induced by the product [3]. Since the human cognitive capacity is a limited resource, increase of cognitive load can affect the users' performance during usage of the product and thereby increase the risk of mistakes occurring.

Unfortunately, poor usability in digital interfaces can be difficult to no-

tice and fix before they actually cause mistakes. One reason for this is the difficulties connected to performing usability-evaluation of human-computer interaction (HCI). The methods used for this evaluation have a history of being expensive, time-consuming and often performed based on poorly documented standards and objectives [4].

One proposed method for studying HCI is through eye-tracking (ET). This technology has during recent years become widely available, as more and more devices now are available on the commercial market. One big advantage with modern ET is that the method is non- to only mildly intrusive and can easily be adapted to a clinical environment, allowing for data acquisition in real-life scenarios [5]. Studies by Goldberg and Kotval [4], Chen et al. [6] and Zagermann et al. [7] within this field have all showed promising results in this methods' ability to distinguish between situations of different cognitive load.

These results suggests that there exist a possibility of using ET as a means of evaluating usability based on quantifiable metrics. This would be preferred as it could possibly simplify the process of usability-evaluation by providing a method that is non-intrusive and can be performed in real-time. The method could also possibly enable objective measuring of cognitive load, as the movements of the eyes in many cases are controlled by unconscious processes [7].

1.1 Objective

The objective of this master thesis was to examine how usability influences cognitive load and eye-movements, with the overall goal of investigating the potential of eye-tracking as a means of evaluating usability. This was done by recording participants' eye-movements when interacting with user-interfaces of varying design quality.

The interfaces were designed to have differences in levels of usability. This was achieved by intentionally disregarding some established design principles. The hypothesis was that this would contribute to a higher cognitive load, which in turn would be objectively reflected in specific eye-tracking metrics.

A visualisation of the hypothesised relationship between the key concepts of the study is presented in figure 1.1.

The results from this study will hopefully contribute to the knowledge regarding eye-tracking as a method of evaluating usability and encourage further research within this field.

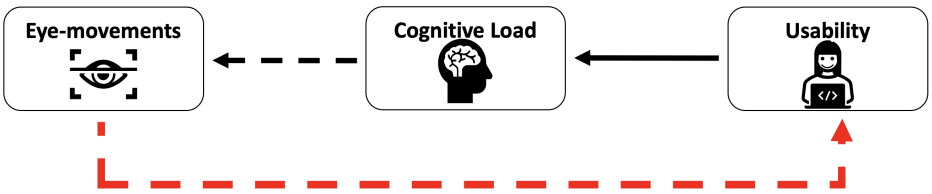


Figure 1.1: Graphical visualisation of the hypothesised relationships between the key concepts *Eye-movements*, *Cognitive Load and Usability*. The dashed arrows represent the connections investigated in the study, while the solid arrows represent assumed connections. The red arrow represents the main objective of the study.

1.2 Research questions

The objective can be expressed as the following main research question and a number of sub-questions

- **Can eye-tracking metrics be used as a method to evaluate usability?**
  - Are eye-tracking metrics affected by the design of a user-interface?
  - Can flaws related to specific design principles be identified using eye-tracking metrics?
  - Can eye-tracking metrics be used as an indicator of cognitive load?

1.3 Disposition of report

This master thesis is divided into six chapters based on the commonly used IMRAD-format, consisting of introduction, method, results and discussion. The format was modified by adding two more chapters, theory and conclusion, to better encompass the scope of the master thesis.

The first chapter gives the reader an introduction to the background, objective and problem statements of the thesis. It is also intended to attract the interest of the reader.

The second chapter presents the necessary theoretical background the thesis is based upon.

Chapter three presents the method used to conduct the experiments and gather necessary data.

The results generated from the data are presented in chapter four and discussed in chapter five.

Finally, chapter six contains a conclusion of the results with regard the research questions and the objective of the report.

Chapter 2
Theory

2.1 Eye movements & eye-tracking

The human visual system is perhaps the most important sense for the everyday activities of humans. Through a complex biological structure, this system translates light into a visual perception of the surroundings. Here, the eyes acts as the outer receptor, responsible for focus as well as direction of our gaze, and is in constant communication with the visual cortex, which interprets the scene and guides our attention [8].

The movement of the eyes is generally divided into two types, fast and slow movements. The fast movements, called *saccades*, are jerking, fast movements that are used to locate points of fixation and interesting targets by searching the visual field. In between saccades, the gaze is focused momentarily in so-called *fixations*, where the most important point in the visual image is focused onto the fovea to achieve optimal sharpness [8]. During these fixations, information about the target of the gaze is processed and interpreted. The interpretation of the scene is then used to decide the subsequent movement of the gaze [5, 9, 10]. The slow movements ensure that focus is maintained by compensating for movement or rotation of the head and movement of the target itself through the visual field [8, 11].

Eye-tracking is a commonly used method for recording eye motion as well as gaze location over time or during a specific task. The method originate from the early 19th century when the physiological connection between the eyes and the nervous system was first defined by Charles Bell in 1823, in a study that connected eye motion to neurological and cognitive processes. Early methods for eye-tracking included mechanical linking of pens to the cornea of the eye, and reflection of light from the cornea to a film of paper [12].

In modern day eye-tracking methods, the gaze is tracked using infrared light illuminating one or both eyes. The light generates a series of reflections, called Purkinje images from the different layers within the eye

[13]. The reflection is then compared with the location of the center of the pupil in order to compute vectors that relate gaze direction to positions in a 2D or 3D-environment, such as on a computer screen *(see figure 2.1)*. This enables the system to track temporal and spatial data, which in turn is used to identify and characterise fixations and saccades. Additionally, some eye-trackers also have the capability to monitor pupil diameter and blinks [9].

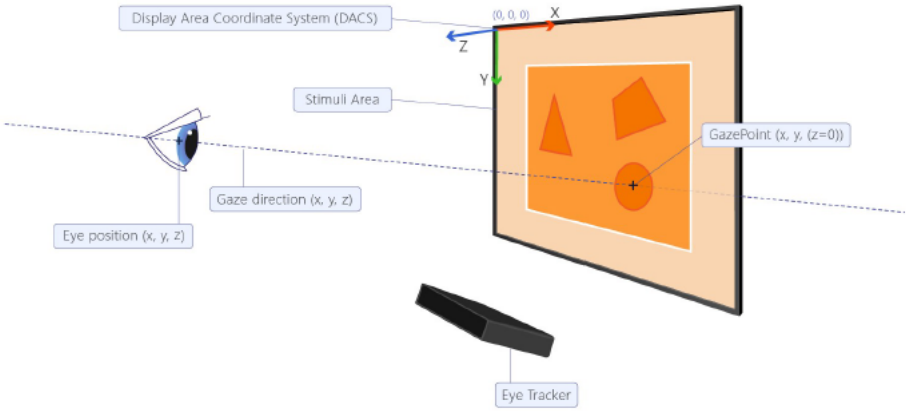


Figure 2.1: Figure showing typical non-invasive eye-tracking setup for tracking gaze on a computer screen. **Image source:** Tobii Pro AB (2014). Tobii Pro Lab User Manual (v1.152.1). Tobii Pro AB, Danderyd, Sweden [14].

2.2 Human computer-interaction

Human computer-interaction (HCI) is the study of how 'computer technology', such as computers with screens and mobile phones, influence human work and activities. It is focused on the interfaces allowing for interaction between the human user and a computer. One discipline of HCI relates to the design of these computer interfaces and is focused on how to make them as easy and pleasant to use as possible. One of the key aspects of this design discipline is the notion of usability, which is often defined in terms of efficiency, effectiveness and satisfaction [15].

2.2.1 Usability

The ISO-definition of usability is *"the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"* [16]. Generally, usability embraces the goals and expectations from the user's perspective during interaction with a product, with the goal to optimise this interaction. For example, aspects such as how easy, enjoyable or effective a product is experienced during usage in their intended environment is key in usability.

One method used to simplify the process of this optimisation is to break down usability in six goals [2]:

- **Effective to use** (effectiveness)
- **Efficient to use** (efficiency)
- **Safe to use** (safety)
- **Having good utility** (utility)
- **Easy to learn** (learnability)
- **Easy to remember how to use** (memorability)

These goals are then typically formulated as detailed questions to facilitate the work for interaction designers, for example:

- *"Does the product provide an appropriate set of functions that will enable users to carry out all of their tasks in the way they want to do them?"* (utility)
- *"Is it possible for the user to work out how to use the product by exploring the interface and trying certain actions? How hard will it be to learn the whole set of functions in this way?"* (learnability)

These questions provides the interaction designer with the possibility of assessing various aspects by concrete examples related to user experience and interaction with the product. This in turn can facilitate formulating specific criterion the product should fulfil and thereby help in ensuring that the end-product has good usability [17].

2.2.2 Design principles

A helpful tool when creating products with a good usability is correct application of design principles. These are fundamental concepts has been formed by a combination of knowledge of professionals from many fields, for example behavioural science, ergonomics and design. There are many examples of established design principles that are used by designers all over the world [18]. Some of these are Jakob Nielsen's *Usability Heuristics*, Donald Norman's *Fundamental Principles of Interaction* and the *Gestalt Theory of visual conception.* Two of these fundamental design principles, which were essential to this master thesis, are elaborated below.

The principle of feedback

Whenever we interact with a product, we need to understand how it works, what it does and what operations are possible. According to design researcher Donald Norman, this can be achieved by appropriate application of his *Fundamental Principles of Interaction* [19]. One of these is the principle of feedback, which means that the results of an action should be communicated. It is a well-known concept from the science of control and information theory. An example of a situation that is a result of a lack of feedback is whenever someone repeatedly pushes the 'Up' button at an elevator. What is lacking in this situation is the feedback that the system is working on your request [19].

The implementation of relevant feedback is a vital part of designing, for example, a web site. One example is the hover event, which can be a helpful feedback tool for a user with a mouse. If the appearance of for example a button or a picture changes when the cursor hovers over it, this can alert the user that the object can be interacted with [20].

Another very important aspect of interface design related to feedback is the responsiveness of the system. A responsive system always keeps the user informed by providing feedback about what has been done and what is happening. A lack of responsiveness can cause confusion, which in turn can be both frustrating and annoying. Research shows that responsiveness is critical to user satisfaction and productivity [21]. An example of poor responsiveness is delayed feedback for button presses. According to the *Time Deadlines for Human-Computer Interaction*, the perception of cause and effect is broken if an action takes longer than 0,1 seconds to show a response. Any action that takes longer than that should display a 'busy indicator'. In the case of pressing a button, the function of a button can take longer, provided that it is shown that the pressing was registered within the deadline [21].

In a study performed in 2009 by Szameitat et al. [22], the empirical effects of sporadic brief delays in HCI were investigated. They used delays with an average duration of 1.6 seconds and their results showed that this delay had a negative effect on for example work productivity, work satisfaction and performance [22].

The principle of proximity and functional grouping

In the early twentieth century, the *Gestalt principles of visual conception* were created by German psychologists. According to these principles, our visual system automatically imposes structure, meaning that we perceive whole figures and objects rather then disconnected areas. These principles have since been used as a basis for guidelines for graphic design and user-interface design [21].

One example of a Gestalt principle that is used frequently in interface design is the principle of proximity. It states that the relative distance between objects decide whether they are organised into groups or not. Objects that are placed closer to each other in relation to other objects appear to be grouped *(see figure 2.2)* [21].

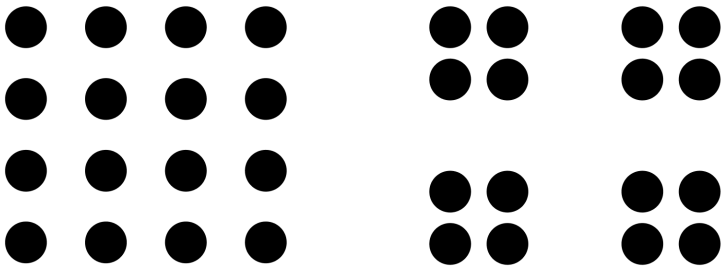


Figure 2.2: A figure showing an example of the principle of proximity. The circles to the left are perceived as one big group while the circles to the right are perceived as four smaller groups.

The principle of proximity is a widespread concept that has been used in various settings in the world of design. It is often combined with the idea of functional grouping when designing a user interface. This is the concept of placing elements that have similar characteristics or functions together in order to simplify for the user. One example is the placement of buttons in a text editor. In this case, sets of buttons with related functions, such as style and alignment, are usually placed together (*see figure 2.3*). This makes it easier for a user to navigate in the user interface. A button placement without functional grouping is likely to be more challenging to

use since it does not provide any clues to the user about where to find the desired button [23].

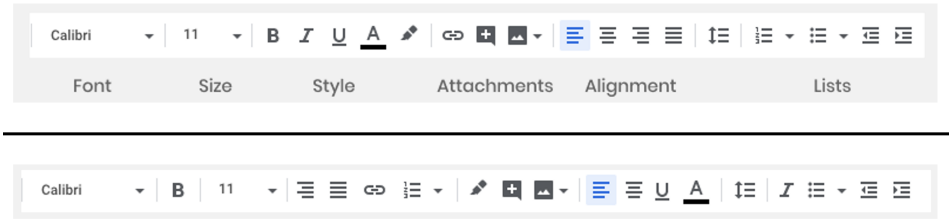


Figure 2.3: Examples of the same text editor both with functional grouping of button (above) and without functional grouping of buttons (below). **Image Source:** Vaz T. Functional Groups — How can other areas of study help usexplain the grouping of elements in design?. Prototypr.io. 2021 Jan 7 [23].

2.3 Cognitive load

2.3.1 Cognitive load theory

Cognitive load theory (CLT) was originally developed by John Sweller in the 1980s. It is based on the assumption that storage and processing of information is performed by two independent systems, the working memory and the long-term memory. The working memory deals with information processing and the long-term memory is used to store information. CLT also assumes that the working memory has a limited capacity and that processing information uses a certain proportion of that capacity [24].

According to CLT, cognitive load can be be divided into three types [24]

- **Intrinsic load** - relates to the complexity of the task
- **Extraneous load** - relates to how the task is presented
- **Germane load** - relates to transferring information from the working memory to the long-term memory

These three load types are additive, meaning that if for example the extraneous load is lowered, working memory capacity to deal with other load becomes available. [3].

2.3.2 Cognitive load, usability and interface design

Even though CLT was originally intended to be used in the field of education, the theory of the limited capacity of the working memory also applies to usability and interface design [25]. In this context, cognitive load imposed by a user interface can be defined as *"...the amount of mental resources that is required to operate the system."* [26]. When designing an interface, the goal should be to minimise the extraneous cognitive load, since it takes up mental resources while not helping the user to understand the content. This results in a higher working memory capacity available to deal with the intrinsic load [26]. When a user has to pause and think while, for example, browsing a web site, their working memory is loaded. This applies even if the pause only lasts for an instant. Examples of questions that can cause these pauses are *"Is this clickable?"* or *"Where is the home button?"* [25]. Therefore, interface design can effect the cognitive load. The extraneous load can be higher with a software design that is suboptimal according to traditional usability goals and a lower extraneous load can thereby create a more usable system. [2, 3].

2.3.3 Measuring cognitive load

Since cognitive load basically is a theoretical construct, it cannot be measured directly. Instead, different measurable parameters assumed to be indicators of cognitive load are used. There are mainly four aspects of these measurable indicators that are used to measure the cognitive load of a task [27].

- Subjective feeling of effort
- Task performance
- Physiological arousal
- Task characteristics

The subjective feeling of effort can be measured using subjective rating scales. The most common procedure is that the test subject first performs a task and directly afterward completes a survey where they rate the level of effort that the task demanded. Despite the perceived risk of personal bias influencing the results, subjective measurements have been found very useful in many domains [28]. There are many established rating scales which have been proven as valid measurements of cognitive load in various studies. One of the most widely used is the NASA Task Load Index (NASA-TLX) [27].

The task performance can be evaluated using different performance measurements. One way to do this is to directly measure how well the task was performed by measuring parameters such as the number of errors committed or the time consumption of the task. In order for this method to be able to differentiate between tasks with different cognitive loads, the difference needs to be on an appropriate level (*see figure 2.4*). If the cognitive load of the tasks are too similar or if the cognitive load of both tasks are either too low or too high, the task performance will not be affected by the differences. These scenarios occur if the tasks to be compared are both in either area A or area C in figure 2.4 [27].

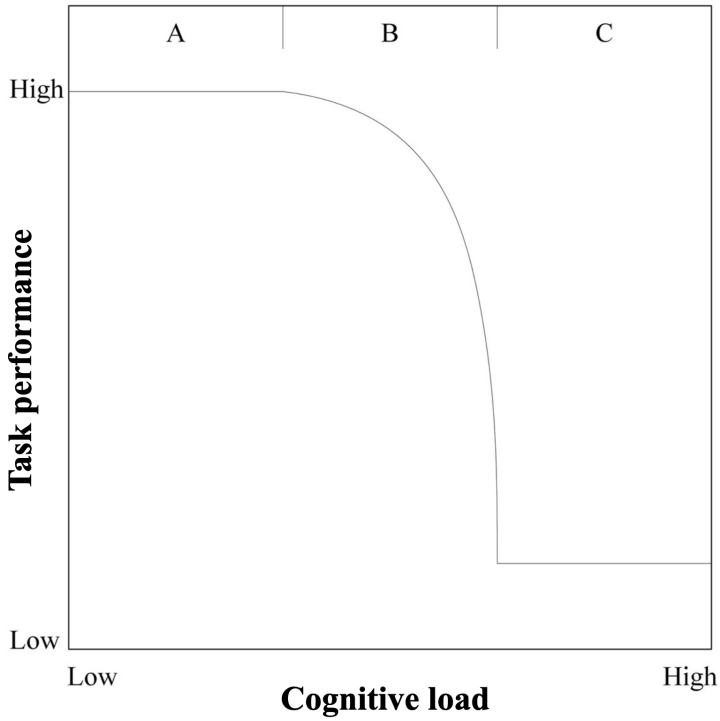


Figure 2.4: A hypothetical relationship between cognitive load and task performance. **Image Source:** Chen S. The Construct of Cognitive Load in Interpreting and its Measurement. Perspectives. 2017 Jan 31; 25(4):640-657 [27].

The physiological arousal that is caused by cognitive load can be measured by analysing many different parameters of the human body. Some examples that have been used in several studies are heart rate, respiration rate and neuroimaging techniques like functional Magnetic Resonance Imaging (MRI). Although there is still no consensus regarding which parameters that best reflect the mental load, recent developments and increased avail-

ability of eye-tracking devices has made it one of the most cost-effective methods for measuring mental workload [24].

The task characteristics can be analysed in order to make a theoretical assessment of the cognitive load of a task. This method can only provide an estimate of the input dimension of the cognitive load. These analytical measures are often provided by experts or derived from task analysis based on current knowledge about the task [27].

NASA-TLX

One of the most commonly used methods to evaluate task load based on subjective measurements is the NASA-TLX questionnaire. It consists of a set of six rating scales of different aspects of workload. These aspects are *mental demand, physical demand, temporal demand, performance, effort* and *frustration* [29]. Various studies have shown that the method performs consistently well regarding for example reliability and user acceptance [28]. However, it cannot be used to obtain real-time information and it also might not be applicable to evaluate unconscious and automatic behaviour of the users [30].

20 years after the NASA-TLX was presented, the research team behind it conducted a follow-up study where the impact and applications of the questionnaire were analysed [31]. They found that various modifications of the original application had been used in other studies. One example was to add, remove or modify the questions in order to increase the relevance to specific studies. They concluded that although modification of the questionnaire can be an excellent strategy to improve its relevance, it is required that the reliability of the modified questions are established before usage [31].

Other common modifications to the NASA-TLX relates to the interpretation of the generated answers. The original method includes a weighting process where the participants rate which of the six measurements that are more relevant to the workload. In a common modification known as the RAW TLX, this weighting process is excluded completely and the overall results are instead generated by simply calculating the average scores. Several studies that had compared this modification to the original version were analysed. It was found that the RAW TLX was either more sensitive, less sensitive or equally sensitive. This indicates that both versions can be used with equal confidence [31].

Another common variation of the NASA TLX is to analyse each subscale individually. This method has proven to be a good way to pinpoint the source of a workload or performance issue [31].

2.4 Experiment design

2.4.1 Randomisation in experimental design

When conducting experimental research, a major reason why causal relations can be found is the usage of randomisation. In an ideal experimental design, the investigator can fully control the experimental conditions to be compared while the other factors are kept the same. One way to achieve this is through randomisation. In a well-designed experiment, factors such as assignment of experiment conditions and order of scenarios needs to be randomised. Randomisation of the order of scenarios can cancel out the potential errors introduced by differences in the scenarios. One commonly used practise in research today is to use software-driven randomisation [32].

2.4.2 Within-group and between group design

When performing an experiment with human participants that investigate the effect of different conditions, there are mainly two ways to distribute the participants, a within-group design and a between-group design (*see figure 2.5*). In a within-group design, each participant is exposed to all conditions. In a between-group design, the participants are instead divided into groups and then each group tests one of the conditions. These approaches have advantages and disadvantages that are opposite to each other [32].

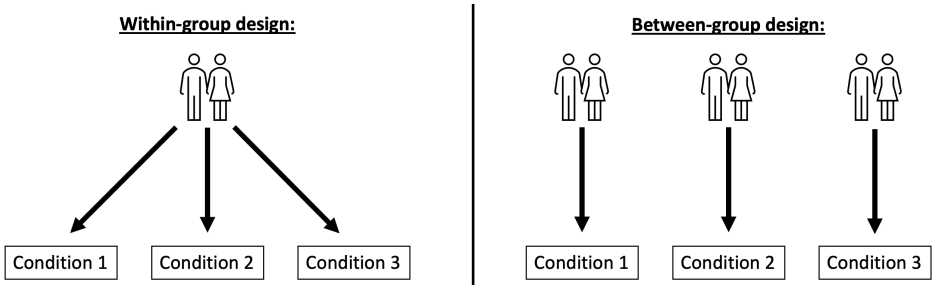


Figure 2.5: An schematic representation of the difference between a within-group design (to the left) and a between-group design (to the right).

Within-group design

As mentioned, an advantage with the within-group design is that it requires a smaller sample size. Since the performance of the same participants under different conditions is analysed, the impact of individual differences is effectively isolated.

The two major disadvantages with a within-group design are the influences of the learning effect and fatigue. Since the participants perform the same type of task under different conditions, they are likely to learn from their experience. This may increase performance over time. The participants may also become tired or bored when they repeat similar tasks multiple times, causing fatigue which may instead decrease performance over time. Both the learning effect and fatigue can be controlled by changing the order that each participant performs the different conditions. One approach is complete randomisation of the order for each participant [32].

Between-group design

The between-group design is a cleaner design from a statistical perspective. Since the participants only test one of the conditions, they are not affected by the learning effect. It is also easier to control other factors that affect a participant, such as fatigue and frustration, since each individual experiment is shorter.

On the other hand, a between-group design is also impacted greatly by individual differences between participants, since one group is compared with another group. It is therefore harder to detect significant differences when using a between-group design. Another disadvantage is that it requires a higher number of participants compared to a within-group design. The difference relates to the number of conditions. If for example an experiment with three conditions is performed, a between-group design needs at least three times more participants compared to a within-group design to achieve the same statistical ground [32].

2.4.3 Random and systematic errors

Whenever an experiment is conducted, there is a risk of errors affecting the results. This risk is especially high when the experiment investigates human behaviour. There are two types of errors, random errors and systematic errors, also called 'biases'. Random errors occur by chance and has no correlation to the actual value. There is no way to eliminate random errors, but their impact can be effectively reduced by enlarging the sample size. Contrastingly, systematic errors always push the observed value away from the actual value in the same direction. Measures must therefore be taken to reduce these types of errors. Some of the major sources of systematic error are [32]:

- Measurement instruments
- Experimental procedures
- Experimental environment

If a measurement instrument used is not accurate or not configured correctly, it may cause systematic error. Ways to prevent this is to use extensively tested and reliable instruments and to examine the instruments used before each session [32].

Experimental procedures that are unclear or inappropriate may cause biases. Instructions that the participants receive needs to be presented in a clear way so that no further explanation is needed. Additionally, it is important that the instructions do not change between tests. If participants are instructed to complete one task *"as fast as possible"* and another task *"in their own pace"*, it may affect the outcome. Two ways to ensure consistency is to present instructions on a written document and to decide in advance if interventions are allowed [32]. Other aspects of the experimental procedure that can cause biases in a within-group design are, as mentioned, learning rate and fatigue. A way to reduce the impact of the learning effect is to provide information and training time to the participants so that they can get acquainted with the task before the experiment starts. Since the learning effect is highest during the initial stages and gradually decrease with further practise, this can help to limit these effects (*see figure 2.6*). A way to reduce the impact of fatigue is to make sure that the test is not to time consuming and to give the participants the opportunity to take breaks if necessary. It is generally suggested that a single experiment session should not last longer than 60-90 minutes [32].

The environmental factors than can cause biases can be divided into two categories; physiological factors and social factors. Examples of some

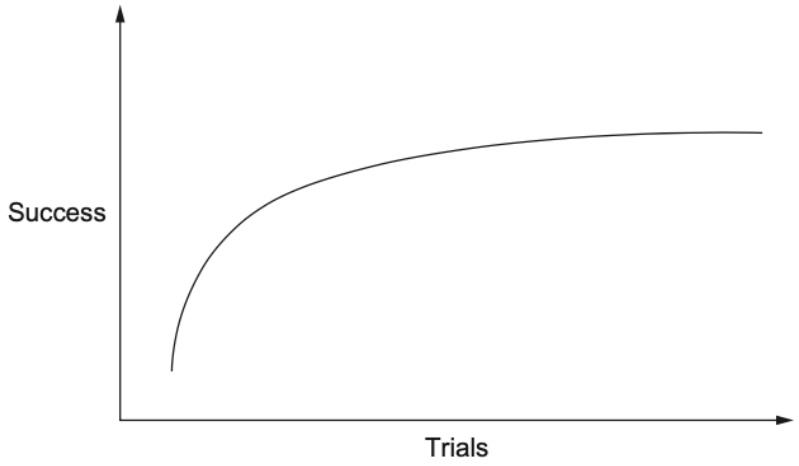


Figure 2.6: A typical learning curve. **Image Source:** Lazar J, Feng JH, Hochheiser H. Research Methods in Human-Computer Interaction. Second Edition. Cambridge, MA, United States:Morgan Kaufmann Publishers; 2017 [32].

physiological factors are noise, lighting, vibrations and temperature. These factors can be controlled by performing all the experiments in the same location, and to make sure that the room is quiet, the lighting is appropriate and that the participants can sit comfortably. It is also important to remove distractions by for example making sure that the room is tidy. An example of a social factor is that a participant having someone looking over their shoulder may perform differently. The participants should therefore be seated alone if possible. It is also important that measures are taken so that participants feel calm and relaxed during the test, by for example reassuring them that it is the interface, not themselves, that is tested [32].

2.5 Previous research

The structure of this study was in many ways similar to a study conducted in 1998 by Goldberg and Kotval [4]. In this study, the authors studied differences in eye movements between two interfaces, one deemed as 'good' according to design principles and one deemed as 'poor'. The distinction between these was based upon the grouping of buttons used in each interface. In the good interface the buttons were grouped based on functionality, while in the poor design, the buttons were randomly grouped. The task given to the participants was to localise and identify a specific button in the interface, while eye-tracking data was recorded. The study found that

eye-tracking had the potential of supplying *"..easily comparable quantitative metrics for objective design iteration."* and allowed estimation of the influence the interface design had on the users' search strategies when using the interface [4].

In a publication from 2003, Goldberg, together with Wichansky, encourage the future development of standards and protocol for the evaluation of usability utilising eye-tracking [5]. The vision here, from the perspective of a usability engineer, is a testing tool that evaluates software with regard to a certain usability-threshold. If the software pass the threshold, it is deemed 'usable', while in the case of it not passing, feedback regarding certain parameters based on the eye-tracking data is given [5].

The measurements used for analysing and how they were interpreted has varied in this field of research. Goldberg and Kotval used a wide variety of metrics to analyse the data from their study. From complex ones, such as convex hull area, transition matrices and spatial density, to simpler ones such as number of fixations and fixation duration. Through their study Goldberg and Kotval found that the 'good' and 'poor' interface could be distinguished by the following metrics; scanpath amplitude, search area, spatial density, transition matrices, number of saccades and number of fixations [4].

Studies within the same subject but with a different approach have been performed since the work of Goldberg and Kotval. A study in 2009 by Chen et al., examined eight eye-based metrics (blink latency, blink rate, mean pupil size, pupil size deviation, fixation duration, fixation rate, saccade amplitude and saccade velocity) and perceived cognitive load during interaction with a computer-based application designed for basketball players [6]. The results from the study showed that all metrics were able to discriminate between two different levels of cognitive load, with saccade velocity and saccade amplitude being notably highly discriminatory [6].

In more recent years, research by Zagermann et al. have both encouraged the usage of eye-tracking as a tool for measuring cognitive load as well as making efforts to construct a model describing the link between eye-based metrics and cognitive load [30]. They suggest number and duration of fixations, amplitude, angle and velocity of saccades, pupil dilation as well as blink rate and blink velocity as eye-based indicators of cognitive load [30]. In a follow-up study from 2018, Zagermann et al. performed a comparison of three tasks with increasingly demanding search processes based on shapes and colors [7]. To evaluate cognitive load for the different tasks, NASA-TLX questionnaires were used. From the results generated in this study, Zagermann et al. conclude that, similar to what both Goldberg and Kotval and Chen et al. found, that some of these eye-based metrics can be used as indicators of cognitive load. The findings showed statistically sig-

nificant differences between the three tasks for fixation rate, saccade rate, blink rate and pupil dilation, where the authors' suggest blink rate and pupil dilation as identifiers of changes in cognitive load and fixation- and saccade rate as indicators of the extent of cognitive load [7].

The above mentioned studies have all concluded that eye-based metrics can be used to distinguish between situations related to varying cognitive load *(see table 2.1 for a compilation of their findings)*. However, as the main objective and characteristics of the tasks within the studies vary, the effect of increased cognitive load might also vary. This finding suggests that the nature of the study are important to consider when analysing and interpreting data, and that conclusions should not be drawn on the effect of one metric alone. Rather, conclusions should be drawn based on the interplay of several metrics, to better encompass the full nature of the task by providing complementary information [6].

Study / **Parameter**	Goldberg and Kotval [4]	Chen et al. [6]	Zagermann et al. [7]	Zagermann et al. [30]*
Scanpath Amplitude	Yes (↑)	-	-	-
Scanpath Duration	No	-	-	-
Number of Saccades	Yes (↑)	-	-	-
Saccade Amplitude	No	Yes (↓)	-	↑
Saccade Rate	-	-	Yes (↑)	-
Saccade Velocity	-	Yes (↓)	-	↑
Number of Fixations	Yes (↑)	-	-	-
Fixation Duration	No	Yes (↑)	-	↑
Fixation Rate	-	Yes (↓)	Yes (↑)	↓
Blink Latency	-	Yes (↑)	-	↑
Blink Rate	-	Yes (↓)	No†	↓
Mean Pupil Size	-	Yes (↑)	-	-

Table 2.1: Table showing the effect on different eye-based metrics with regard to increased cognitive load based on results from three experimental and one theoretical study. 'Yes/No' indicates whether a statistically significant difference was found or not and the arrow indicates if the metrics increased (↑) or decreased (↓) with increased cognitive load. '-' denotes that this metric was not included in the study.
*This study was purely theoretical, based on data generated from experiments related to the subject.
†Statistically significant difference was initially found. However, post-hoc analysis applying Bonferroni correction revealed no significant difference.

Chapter 3
Method

3.1 Summary of experiment

Eye-tracking data was collected during experiments in which the participants interacted with three different interface-prototypes. In two of the prototypes, a fundamental design principle was intentionally ignored to increase the experienced cognitive load during interaction with the prototype and by doing so also decreasing usability. The remaining prototype, where fundamental design principles were followed, acted as a point of reference.

The recorded eye-tracking data, along with a subjective evaluation of cognitive load and sound recordings from the experiment, was then analysed based on a number of pre-defined metrics. The results were then examined through statistical analysis to see whether the results could be used to distinguish between the three prototypes.

3.2 Interface prototypes

3.2.1 The prototype task

The task that was to be performed in the prototypes was a decision that was made early in the project. The main goal was to create a 'realistic' scenario, since previous studies performed in the same area had been very restricted or unrealistic. One example is the study by Goldberg and Kotval [4], where the task was to localise one specific button at a time and thereby heavily restricting the participants. Another example is the study by Zagermann et. al [7], where the task was related to identifying shapes and colors which has no realistic application. On the other hand, it was important to still have some control over the participants' actions in order to allow for comparisons between both the prototypes and the participants. The balance between restriction and control was achieved by giving the participants the ability to choose the order in which they performed the actions of each task, while keeping the end result of each task constant for all participants.

Another idea was to simulate a quite simple task that most people are familiar with in order to decrease the influence of the learning effect (*see section 2.4.3*).

The end result became a task where the user was to create and manipulate a 'post-card', consisting of a picture and a text box containing a standard greeting (*see figure 3.1*). The three pictures that were included in the prototypes depicted a football player, a beach and skiers on a snowy mountain. Some examples of functions implemented in the prototype was the ability to change the color and the font of the text, change the position of the text box and apply a filter to the picture. The functions were applied by clicking a designated button with an icon representing each function respectively. Included functions and icons were intended to be familiar from commonly used interfaces and some sources of inspiration were Microsoft Word, Microsoft Paint and Instagram.

Figure 3.1: An example of a 'post-card' from the prototypes.

3.2.2 The prototype design

In this project, prototypes that differ in interface design were created. After discussion with the supervisor in the early stages of the project, it was decided that three different prototypes were to be created. One with a 'good' design that would function as a reference that the others could be compared with (*see figure 3.2*). The other two interfaces were designed to be 'worse' in some way, by intentionally breaking aspects of a fundamental design principle, resulting in interfaces with poorer usability. The two design principles that were broken were the principle of proximity and the principle

of feedback. These principles were chosen because they of their relevance in interface design and because they were assumed to have a direct effect on the eye-movements. They were also deemed fairly simple to implement into the prototypes.

The overall goal of these prototypes was that the intrinsic load, the load related to the complexity of the task, would be the same for all three prototypes while the extraneous load, the load related to how the task is presented, would be higher in the two 'worse' prototypes due the changes to the interface design. However, the changes in the design were kept fairly small, in order not to loose the aspect of realism in the test.

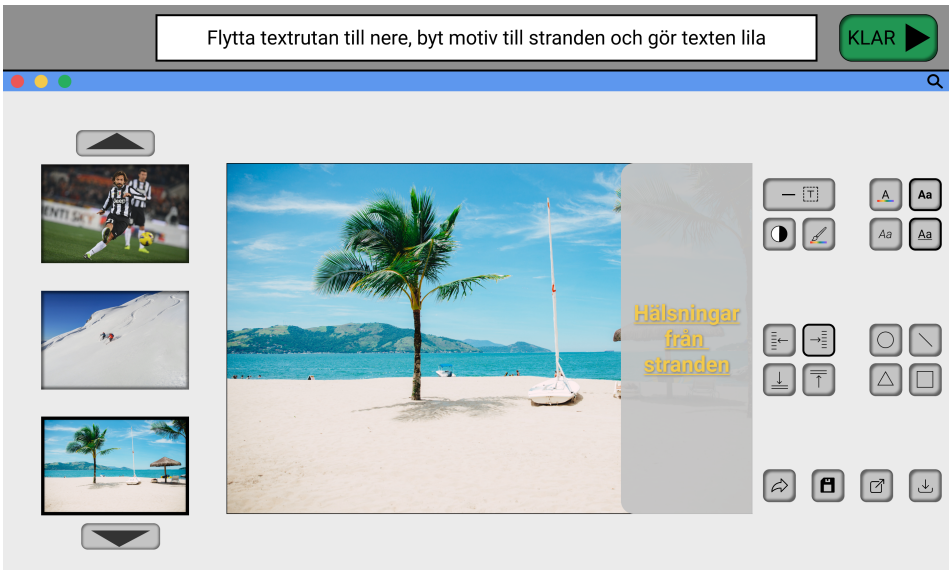


Figure 3.2: An example of a full view from the 'good' prototype.

Sorting prototype

In the 'sorting' prototype, the grouping of the buttons based on function was removed in order to complicate the search process (*see figure 3.3*). This was a violation of the principle of proximity from the Gestalt Theory (*see section 2.2.2*). Instead of a complete randomisation of the button placement, they were sorted in rows based on their function. This is one example of a measure taken to maintain the realistic aspect of the test.

The idea with this change was that the extraneous cognitive load would increase due to that the search process was intentionally complicated.

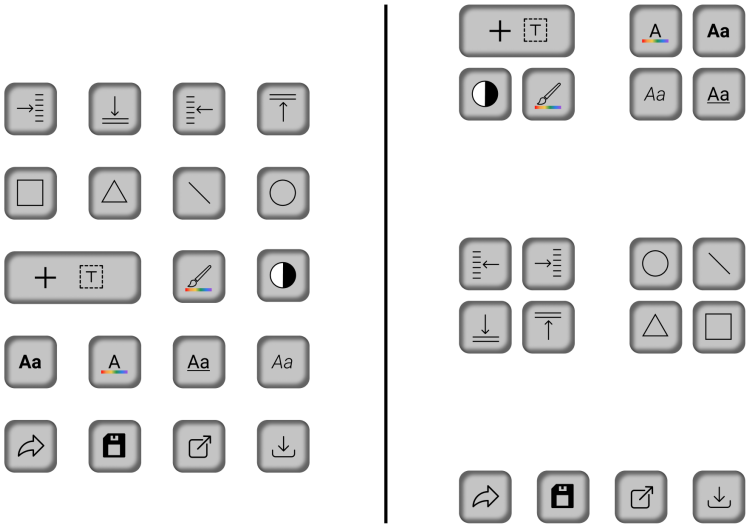


Figure 3.3: The difference in the button placements between the 'sorting' prototype (left) and the two other prototypes (right).

Feedback prototype

In the 'feedback' prototype, less information was presented to the user when interacting with the system, violating the design principle of feedback (*see section 2.2.2*). The first difference was related to direct visual feedback. In the first two designs, a hover effect was added to each button, as a way to inform the user that is was clickable. This was removed in the third design.

The second difference related to responsiveness. All three designs contained a built in delay of 1.6 seconds that was randomly distributed to some of the actions. The length of this delay was based on the research by Szameitat et al. (*see section 2.2.2*). The difference was that the 'good' and 'sorting' prototype had an implemented 'busy indicator' in the form of an hourglass (*see figure 3.4*) that informed the user the action had been registered. The 'feedback' prototype instead presented no information to the user about the delay.

With these changes the extraneous cognitive load is expected to increase due to confusion and possible frustration occurring as an effect of the changes.

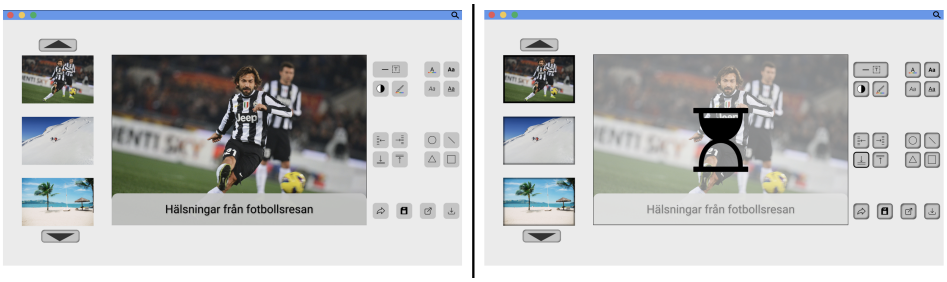


Figure 3.4: The difference in responsiveness between the prototypes. The 'feedback' prototype presented no 'busy indicator' (left) while the two other did (right).

3.2.3 Prototype creation

The prototypes was created using the interface design software Figma (https://www.figma.com). Fistly, a draft of the prototype was created in order to test out the functions. After analysis and discussion with the supervisors, this draft became a proof of concept for the prototypes. The draft was then reworked into the final three designs in an iterative process where design suggestions were sent back and forth between the authors and the supervisors a few times.

All actions to be performed were presented to the user in form of a direct command and the wording for each action was identical between the prototypes, to decrease the possibility of potential bias *(see section 2.4.3)*.

Which actions that was to be performed by the user in each test was decided through randomisation. The delay that was build into the system was also distributed at random, affecting roughly 50% of the actions. All randomisations were made using a randomisation web site. This was a way to remove all human bias in the creation of the test and to ensure a more ideal experimental design (*see section 2.4.1*). This randomisation was extra important since only one prototype was created for each of the three interface designs, meaning that all participants performed the same actions when using for example the prototype with less feedback. This is not deemed to have affected the outcome of the experiment, based on the randomisation and the familiarity and simplicity of the actions.

3.3 Experiment

3.3.1 Participants

In this experiment, the participants were distributed according to a within-group design (*see section 2.4.2*). The main reason for this choice was because of an uncertainty in the number of participants that could be acquired due to the on-going Covid-19 pandemic. Another reason was to eliminate individual differences between the participants from affecting the results.

In total 33 participants, 17 female and 16 male, between the ages of 20 and 29 contributed to the study. All participants were fluent in Swedish, as was needed since the information in the tests was presented in Swedish. Because of issues regarding the performance of the test and calibration of the eye-tracker, data from three participants was disregarded in the final data analysis, resulting in a final data-set containing data from the remaining 30 participants.

To achieve such a homogeneous group as possible, participants were chosen from a restricted age-group. This also ensured that the level of computer-skill was similar between all participants. No discrimination was made based on known visual defects or usage of glasses or lenses as the system used was deemed sturdy enough to handle this without affecting the data significantly.

3.3.2 Apparatus

The test-setup consisted of two computers, an external presenter screen, eye-tracker, eye-tracking software, audio interface, microphone, wireless computer mouse and a divider *(see figure 3.5)*. The ET-software, [Tobii Pro Lab (Version 1.152.30002x64) Danderyd, Sweden: Tobii Pro AB] ran on computer 1 which in turn was connected to the eye-tracker [Tobii Pro Nano Danderyd, Sweden: Tobii Pro AB] as well as to the presenter screen [size=21.5", resolution=1920x1080p]. This computer used screen-extension to present the prototypes on the presenter screen while also allowing for the test leaders to control and monitor the activity in real-time on the internal screen of computer 1. A wireless computer mouse was connected via bluetooth to computer 1, allowing the participant to interact with the prototypes during the tests. The test leaders used the internal trackpad of computer 1 to interact with the used software.

Computer 2 was used for audio recording during the test. This was done by connecting it to a microphone placed on the same surface as the presenter screen, via an audio interface.

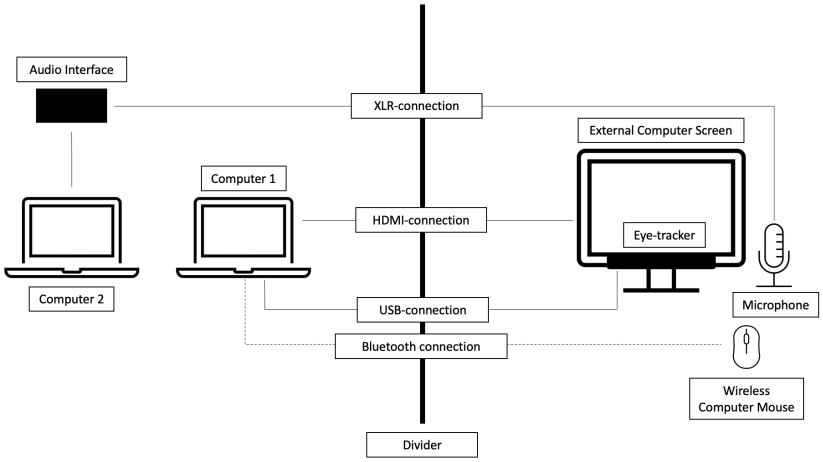


Figure 3.5: Test-setup used during the testing.

3.3.3 Procedure

After arriving at the building in which the tests were conducted, the participant was escorted to the usability-lab, in which the test was performed, by one of the test-leaders. Tests were performed with one participant at a time and the participant was also separated from the test-leaders with a divider. They were also asked to put their cellphone in a 'do not disturb'-mode. This setup removed unnecessary distractions, ensuring that both the physiological and social aspects of the experimental environment were controlled in order to lower the risk of potential biases (*see section 2.4.3*).

Once in the lab, the test-leaders gave a quick introduction to the procedure and aim of the experiment and answered any questions to make sure that the participant felt comfortable. The participant was then asked to read and sign a form of consent *(see Appendix B)*.

After this, the participant was instructed to take a seat in front of the external screen with the attached eye-tracker and adjust the chair so that they sat comfortably. The participant was then given a introduction to both the general function of the prototypes and the function of each individual button. These instructions were intended to be thorough. The instructions were also presented as written instructions in a slideshow, which ensured that the information-presentation was consistent between all participants. This was done in order to reduce the bias caused by a unclear or inconsistent experimental procedure *(see section 2.4.3)*. After the instructions, the eye-tracker was calibrated. If necessary, the participant was asked before the calibration to move the chair so that there was a distance of 65±5 cen-

timeters between the face of the participant and the eye-tracker and to angle the screen so that their face was centralised by the eye-tracker. The calibration was then performed using the built-in function in Tobii Pro Lab, using five points of calibration and four points of validation. The calibration was deemed successful if the validation accuracy and precision did not exceed 1.5 degrees. Both the calibration of the eye-tracker and the alignment of the participants face was done to lower the risk of biases from measurement instruments (*see section 2.4.3*).

The test of the first prototype was then initiated. Each participant performed one test per prototype, resulting in three tests per participant. Each test consisted of five prompts, were each prompt in turn consisted of three actions needed to fulfil the prompt *(see figure 3.6)*. For example, one prompt could be: *"Switch picture to skiing, make the text green and move the textbox to the right-hand side"*. These actions could be performed in any order. When the actions needed had been taken, the participant could move on to the next prompt. The five prompts within each prototype-specific test did not vary between participants, but did vary between the different prototype-specific tests.

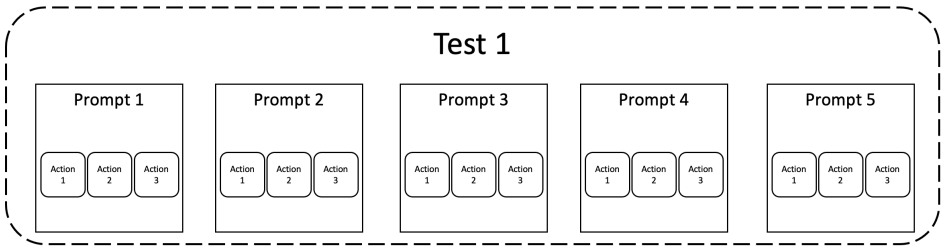


Figure 3.6: An overview of the hierarchy of a test.

After completing all the prompts the recording was stopped and the participant was instructed to fill out the NASA-TLX questionnaire *(see Appendix C)*. The procedure (calibration, test and evaluation) was then repeated two more times for the two other prototypes (*see figure 3.7*).

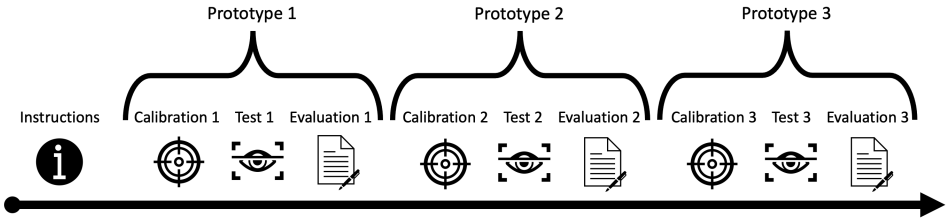


Figure 3.7: The overall timeline for the experiment

The order of the three prototypes ('good', 'sorting' and 'feedback') for each participant was randomly decided by the test-leaders. This was done to avoid the structure of the experiment becoming a factor influencing the results, which is otherwise a risk when conducting an experiment with a within-group design *(see section 2.4.2)*.

Sound was recorded from the start of each test until the end of the test to enable counting of the number of mouse clicks performed. Recorded ET-metrics were exported from Tobii Pro Lab and statistically analysed. To find onset- and offset-timestamps of the tasks from the prototype testing, the ET-recording was manually analysed in the built-in video presenter in Tobii Pro Lab.

3.4 Data analysis

From the recorded ET-data, sound recording and evaluation-form, a series of metrics were assessed. The assessed metrics derived from these data were:

- **Number of fixations** - The total number of fixations performed by the participant during the test.
- **Fixation duration** - The duration during which fixations take place.
- **Fixation rate** - The number of fixations performed per second.
- **Saccade amplitude** - The distance, in degrees, on the screen between two subsequent points of fixation.
- **Task amplitude** - The total distance the gaze travels on the screen from the onset of a prompt until completion of the same prompt.
- **Task duration** - The duration from onset of a prompt until the completion of the same prompt.
- **Excessive mouseclicks** - The total number of mouseclicks performed during the test minus the minimal amount of mouseclicks needed to complete the test.
- **Mental demand** - The mental demand experienced by the participant during the test.
- **Physical demand** - The physical demand experienced by the participant during the test.
- **Temporal demand** - How time-consuming the participant experienced the test to be.
- **Performance** - How successful the participant deemed their performance to be during the test.
- **Effort** - How strenuous the test was experienced to be by the participant.
- **Frustration** - How frustrating the test was experienced to be by the participant.

Statistical comparison between the data from the three prototypes was performed with repeated-measures ANOVA. If a significant difference was found between the prototypes for a specific metric, a post hoc test was

performed with pairwise comparison using Fisher's Least Significant Difference. The statistical analysis was conducted using IBM SPSS Statistics [IBM Corp. Released 2020. IBM SPSS Statistics for Macintosh, Version 27.0. Armonk, NY: IBM Corp].

3.4.1 Eye-tracking metrics

The specific ET-metrics that were to be analysed from the ET-data was decided based on previous research, mainly that of Chen et al. [6] and Zagermann et al. [7], but also that of Goldberg and Kotval [4] *(see section 2.5)*, and the nature of the task. Furthermore, the metrics should not be too complex to facilitate simple usage and structuring of future testing. Metrics that had been used in multiple studies were prioritised to ensure that the generated results could be compared to that of previous research. In this way, differences between the different tasks in previous research and their effects could be taken into consideration.

The final ET-metrics used for evaluation of the tests were *number of fixations, fixation duration, fixation rate, saccade amplitude* and *task amplitude*. Based upon the results from previous studies, a hypothesis for the expected effect of increased cognitive load on each metric was formulated. These hypotheses are presented in italics in the sections below.

Number of fixations

The number of fixations is expected to increase with increased cognitive load. However, this metric might also be highly related to overall test-duration.

According to Goldberg and Kotval, the total number of fixations is related to components required to be processed in a task, but not the depth of the processing. Thus, a large number of fixations would indicate a distracted or complicated process up until decision-making. The study performed by Goldberg and Kotval found that the number of fixations increased by 17% during interaction with their 'Poor' interface compared to their 'Good' interface [4].

Fixation duration

An increase of cognitive load is expected to increase the fixation duration, with the reservation that the difference might not be notable due to that the representation is kept identical between the prototypes.

The duration of a fixation has, according to Zagermann et al., been related to the level of cognitive processing, where higher fixation duration indicates

increased strain on the working memory [30]. Results from a study performed by Chen et al. agreed with this hypothesis as the results indicated a increase of fixation duration as the cognitive load was increased [6].

Worth noticing however, is that Goldberg and Kotval found no statistically significant difference in fixation duration between the two interfaces used in their study. They argue that fixation duration is related to representation, for example how easy it is to correctly interpret the function of a button. A representation that require longer fixation duration is not as meaningful to the user and would because of this result in higher cognitive load. As the representations did not differ between the two interfaces used in their study, the authors were not surprised that the difference with regard to this metric was not statistically significant [4].

Fixation rate

The fixation rate is expected to increase with increased cognitive load. This might however be the opposite way around for a study where the parameter deciding cognitive load is more dependent on attention.

The proposed effect on fixation rate that Zagermann et al. presented in their theoretical study is that fixation rate decrease with increased cognitive load [30]. This claim is based on the results from the experimental study performed by Chen et al., where the observation was that fixation rate significantly decreased with increased cognitive load [6].

Interestingly, this claim stands in contrast to the results generated in a subsequent study by Zagermann et al., where the results instead indicate a increase of fixation rate with increased cognitive load [7]. This contradiction most likely emerge from the different measures taken by the authors to increase cognitive load. Here, the difficulty of the task in the study conducted by Chen et al. was based on the amount of focus and attention required by the participant to succeed with the task, while the difficulty in the study by Zagermann et al. was based around complicating the search-process [6, 7].

This example highlights the importance of interpreting results based on the nature of the task and thereby also the importance of the interpretability of used metrics.

Saccade amplitude

The saccade amplitude is expected to decrease with increased cognitive load.

In the theoretical study by Zagermann et al., it is suggested that the saccade amplitude would increase with a higher cognitive load [30]. On the other

hand, Goldberg and Kotval suggest that the saccade amplitude would instead decrease with a higher cognitive load. This is based on the assumption that a well designed interface will provide more cues to direct the scanning of the user, resulting in longer saccades with fewer interim fixations [4].

Experimental studies of the effect of increased cognitive load on the saccade amplitude show different results. Goldberg and Kotval obtained no significant difference in average saccade amplitude. They came to the conclusion that although the well designed interface resulted in a overall shorter search, the individual saccadic motions were not affected. The study by Chen et al. instead showed a significant decrease in average saccadic amplitude, which supports the theory of Goldberg and Kotval. However, Chen et al. also present examples of other studies that showed no systematic change of the saccade amplitude with regard to increased workload. It is suggested that this is caused by differences in the nature of the tasks in the respective studies [6].

Task amplitude

The task amplitude is expected to increase with increased cognitive load

From the study performed by Goldberg and Kotval the results show that the two interfaces could be separated by the length the gaze travels on the screen from start of a task until the task was completed. This measure is argued to be a practical measure of how extensive the search behaviour is as search processes might have similar duration, but differ in length [4].

Excluded metrics

The metrics *number of saccades, saccade speed* and *saccade rate* were deemed redundant as they were considered to highly correlated with other metrics used in the study (number of fixations, saccade amplitude and fixation rate) [4, 30]. Metrics involving the pupil was excluded due to the difficulty in controlling external parameters that can have a big impact on pupil size, such as lighting and physical surroundings [30]. Finally, as the hardware used in the data collection was unable to detect blinks, metrics related to this, such as blink rate and blink latency, were excluded.

3.4.2 Subjective evaluation

The ratings of the prototypes from the NASA-TLX questionnaire is expected to increase with increased cognitive load.

The method that was chosen to evaluate the cognitive load using a subjective measurement was the well established NASA-TLX questionnaire (*see*

section 2.3.3). All six rating scales were included without being modified despite that some of them, especially 'physical demand', were not entirely applicable. This choice was made since the required effort to ensure the reliability of a NASA-TLX with modified questions was deemed to be outside the scope of the report. All rating scales were structured so that a more 'positive' experience, i.e. lower cognitive load, will result in lower ratings, while high ratings would indicate a less 'positive' experience.

However, small changes were made in the interpretation of the generated data. The common modification known as the RAW TLX was chosen. This was due to the fact that it is simpler to apply. It was also chosen that each subscale would be compared individually in addition to the average score. This was a way to generate as much information as possible from the questionnaire.

3.4.3 Task performance

Two metrics were used during the tests for evaluating task performance as a complement to the subjective evaluation *(see section 2.3.3)*.

Task duration and Excessive mouseclicks

Task duration and Excessive mouseclicks are both expected to increase with increased cognitive load.

Measuring the time needed to complete the tasks in each test act as a key to the rating scale 'temporal demand' on the NASA-TLX questionnaire, as well as an indicator of any errors committed which would prolong the task.

Counting the number of performed mouseclicks and comparing it to the minimum amount needed to complete each test indicate the overall performance experienced by the participant during the test, comparable to the rating scales 'performance', 'effort' and 'frustration' on the NASA-TLX questionnaire. Additionally, it could also indicate the magnitude of errors committed.

3.5 Measures taken due to the COVID-19 pandemic

Due to the ongoing pandemic, the process of data collection was aggravated. In order to minimise potential exposure to and spreading of disease, a series of measures was taken.

The test-setup was structured in a way that enabled the test leaders to maintain distance to the participant during the entire data collection. Additionally, a divider separated the test leaders and the participant during the duration of the test *(see figure 3.5)*. The test leaders wore face-masks

during the entire duration of the data collection. To minimise the exposure of other persons present in the building, the participant was escorted both to and from the test-room out of the building by one of the test-leaders. Surfaces, such as computer screen, computer mouse and computer desk was cleaned with sanitiser between each participant.

To organise the data collection, a bookable calender was used were participants could book a pre-defined time-slot that suited them. This ensured that there was enough time for escorting, cleaning and preparation between each participant. Participants were also chosen from the test-leaders' personal acquaintances. This was done to facilitate possible infection tracing in case of infection among the participants or authors.

Chapter 4
Results

The results from the experiments and the calibration are presented in this chapter in the form of text as well as in tables 4.1, 4.2 and 4.3. For a visual presentation of the experimental results see Appendix A.

4.1 Eye-tracking metrics

Eye-tracking metrics			
	Good	**Sorting**	**Feedback**
Number of fixations	227.93 (±37.85)	244.90 (±49.21)	236.43 (±57.44)
Average fixation duration (ms)	281.77 (±33.72)	263.77 (±28.76) (↓)	277.57 (±30.0)
Fixation rate (Hz)	2.89 (±0.33)	3.07 (±0.34) (↑)	2.93 (±0.3)
Average saccade amplitude (°)	5.49 (±0.65)	5.07 (±0.73) (↓)	5.44 (±0.67)
Average task amplitude (°)	202.81 (±42.13)	201.52 (±49.03)	209.6 (±51.06)

Table 4.1: Summary of results from the ET-metrics. Each cell contains the mean and standard deviation based on metric and prototype. Cells coloured green indicate that a statistically significant difference was found in comparison to the good prototype, while cells coloured red indicate that no difference was found. The arrow in the green coloured cells indicate either a increase (↑) or decrease (↓) for that metric in comparison to the good prototype.

4.1.1 Number of fixations

*The number of fixations was **expected to increase**. According to the results, the number of fixations **did not differ significantly** between **neither** the good prototype and sorting prototype **nor** between the good prototype and feedback prototype.*

No statistically significant difference was found between the prototypes $F(2,58)=1.306$, $p>0.05$ ($p=0.279$).

The number of fixations did increase slightly when comparing the good prototype (M=227.93, SD=37.85) to the sorting prototype (M=244.90, SD=49.21) and the feedback prototype (M=236.43, SD=57.44).

4.1.2 Average fixation duration

*The fixation duration was **expected to increase**. According to the results, the average fixation duration was **significantly shorter** for the **sorting prototype** in comparison to the **good prototype**. There was **no significant difference** between the **feedback prototype** and the **good prototype**.*

Statistically significant difference between the prototypes was found, $F(2,58)=9.297$, $p<0.05$ ($p<0.001$). Post hoc-analysis showed that there was a significant difference when comparing the good prototype to the sorting prototype, ($p<0.001$). No significant difference was found between the good prototype and the feedback prototype ($p=0.363$).

The average fixation duration was significantly shorter for the sorting prototype (M=263.77, SD=28.76) than for the good prototype (M=281.77, SD=33.72). The average fixation duration for the feedback prototype (M=277.57, SD=30.0) was slightly shorter than for the good prototype.

4.1.3 Fixation rate

*The fixation rate was **expected to increase**. According to the results, the fixation rate was **significantly higher** for the **sorting prototype** in comparison to the **good prototype**. There was **no significant difference** between the **feedback prototype** and the **good prototype**.*

Statistically significant difference between the prototypes was found, $F(2,58)=11.001$, $p<0.05$ ($p<0.001$). Post hoc-analysis showed that significant difference was found when comparing the good prototype to the sorting prototype, ($p<0.001$). No significant difference was found between the good prototype and the feedback prototype ($p=0.236$).

The fixation rate was significantly higher for the sorting prototype (M=3.07, SD=0.34) than for the good prototype (M=2.89, SD=0.33). The

feedback prototype (M=2.93, SD=0.3) had a slightly higher fixation rate than the good prototype.

4.1.4 Average saccade amplitude

*The saccade amplitude was **expected to decrease**. According to the results, the average saccade amplitude was **significantly lower** for the **sorting prototype** in comparison to the **good prototype**. There was **no significant difference** between the **good prototype** and the **feedback prototype**.*

Statistically significant difference between the prototypes was found, $F(2,58)=12.895$, $p<0.05$ ($p<0.001$). Post hoc-analysis showed that significant difference was found when comparing the good prototype to the sorting prototype, ($p<0.001$). No significant difference was found between the good prototype and the feedback prototype ($p=0.604$).

The average saccade amplitude was significantly lower when comparing the good prototype (M=5.49, SD=0.65) to the sorting prototype (M=5.07, SD=0.73). The average saccade amplitude of the feedback prototype (M=5.44, SD=0.67) was slightly lower than the good prototype.

4.1.5 Average task amplitude

*The task amplitude was **expected to increase**. According to the results, there was **no significant difference** in average task amplitude **neither** between the good prototype and the sorting prototype **nor** between the good prototype and the feedback prototype.*

No statistically significant difference was found between the three prototypes, $F(2,58)=0.453$, $p>0.05$ ($p=0.638$).

The average task amplitude was similar for all prototypes: good prototype (M=202.81, SD=42.13), sorting prototype (M=201.52, SD=49.03), feedback prototype (M=209.6, SD=51.06).

4.2 Performance metrics

Performance metrics			
	Good	**Sorting**	**Feedback**
Average task duration (s)	13.75 (±2.23)	14.02 (±2.62)	14.06 (±3.67)
Excessive mouseclicks	2.8 (±2.06)	3.77 (±5.81)	4.1 (±6.05)

Table 4.2: Summary of results from the performance metrics. Each cell contains the mean and standard deviation based on metric and prototype. Cells coloured green indicate that a statistically significant difference was found in comparison to the good prototype, while cells coloured red indicate that no difference was found.

The performance metrics were ***expected to increase****. The results from the performance metrics used* ***showed no statistically significant difference*** *between the three prototypes.*

Neither of the metrics related to task performance, excessive mouseclicks and average task duration, was able to distinguish between the prototypes. Generally, the good prototype caused a smaller deviation between participants' performance. However, the best performances related to task duration was seen in interaction with the the feedback prototype and the sorting prototype.

4.3 Subjective evaluation

Subjective evaluation			
	Good	**Sorting**	**Feedback**
Mental demand (NASA-TLX)	4.63 (±3.13)	5.53 (±3.5)	4.67 (±3.54)
Physical demand (NASA-TLX)	2.03 (±1.47)	2.07 (±1.55)	2.17 (±1.72)
Temporal demand (NASA-TLX)	4.2 (±2.82)	4.2 (±3)	4.47 (±3.6)
Performance (NASA-TLX)	3.6 (±3.46)	4.37 (±4.51)	3.8 (±4.25)
Effort (NASA-TLX)	4.4 (±3.43)	4.97 (±4.44)	4.53 (±4.34)
Frustration (NASA-TLX)	5.13 (±4.87)	4.6 (±4.11)	5.33 (±4.85)

Table 4.3: Summary of results from the used subjective evaluation. Each cell contains the mean and standard deviation based on metric and prototype. Cells coloured green indicate that a statistically significant difference was found in comparison to the good prototype, while cells coloured red indicate that no difference was found.

*The ratings from the NASA-TLX questionnaire were **expected to increase**. The results from the subjective evaluation show that the participants' **perceived** cognitive load **did not statistically differ significantly** between the prototypes.*

The analysis of the subjective evaluation showed that none of the three prototypes was distinguished based on any of the six rating scales. The rating scale closest to being able to distinguish any prototype was 'mental demand' where the sorting prototype (M=5.53, SD=3.5) was rated as slightly more demanding than both the good prototype (M=4.63, SD=3.13) and the feedback prototype (M=4.67, SD=3.54). However, this difference was not statistically significant, $F(2,58)=2.284$, $p>0.05$ ($p=0.111$).

4.4 Calibration results

	Mean	**Standard deviation**
Validation accuracy (°)	0.617	0.301
Validation precision (°)	0.516	0.351
Gaze samples (%)	95.47	2.63

Table 4.4: Mean and standard deviation values for validation accuracy and precision as well as proportion of valid gaze samples from the performed calibrations. Calibration was performed using the built-in function in Tobii Pro Lab (Version 1.152.30002x64).

Chapter 5

Discussion

Usability evaluation of human-computer interaction has a history of being expensive, time-consuming and performed based on poorly documented standards and objectives. Compared to previously used methods for usability evaluation, eye-tracking (ET) has the potential to offer a cost-effective method that provides objective and quantifiable measures in real-time. The objective of this master thesis was to examine the influence of usability on cognitive load and eye-movements, with the overall goal of investigating the potential of ET in usability evaluation.

This was done by analysing data gathered from an experiment in which 30 participants interacted with three interface-prototypes designed to have good or poor usability. The differences in usability were achieved by intentionally disregarding some fundamental design principles, namely the principle of feedback and the principles of proximity and functional grouping. By disregarding these principles, the usability of the prototype was impaired, leading to an increase of the cognitive load *(see section 2.3.2)*. The data used to analyse the prototypes were based on subjective evaluation of cognitive load, performance metrics and ET-metrics.

The results show that the ET-metrics were able to distinguish between the reference ('good') prototype and the prototype where the principles of proximity and functional grouping were disregarded ('sorting'). This indicate that the ET-metrics were able to identify differences in design of a user-interface related to this specific design principle.

Neither the results from the subjective evaluation of cognitive load nor performance metrics indicated any significant differences between the three interface-prototypes. These results dispute what was seen from the eye-tracking metrics, since the effects on the eye-movements were expected to be caused by changes in cognitive load. This could be explained by the possible hypothesis that the eye-tracking was able to capture the distinction between perceived and actual cognitive load.

Principles of proximity and functional grouping

The results from the study show that three of the ET-metrics used, fixation duration, saccade amplitude and fixation rate, were able to distinguish between the 'good' prototype and the 'sorting' prototype.

The effect on two of the metrics, fixation rate and saccade amplitude, aligns with what was expected from previous research *(see section 3.4.1)*, i.e. increase of fixation rate and decrease of saccade amplitude with increased cognitive load. Regarding fixation duration, the average duration was significantly shorter for the 'sorting' prototype than for the 'good' prototype, contrary to what was expected based on the previous research.

However, these effects align with expectations based on the difficulties present in the 'sorting' prototype, where the search process was intentionally complicated. Without the clues that a functional grouping of buttons presents, each fixation gives the user less information regarding the location of the desired button or function. This naturally results in more fixations of shorter duration. Omitting the functional grouping of buttons should also lead to less purposeful search, resulting in lower saccade amplitude, which was seen from the results.

Subjective evaluation of cognitive load

The results from the subjective evaluation of the prototypes differ from the expectation that disregarding fundamental design principles would result in an increase of the perceived cognitive load. Although there are some tendencies that the 'sorting' prototype caused a higher cognitive load, especially when looking at mental demand, no statistically significant differences between the prototypes were found.

Herein lies a problematic discrepancy between the results from the ET-metrics and the subjective evaluation, negatively affecting the certainty in the conclusions able to be drawn from the results. The proposed overall relation between usability, cognitive load and eye-movements is based around the idea that changes in cognitive load is reflected in the eye-movements. While not rejecting that this might be the case, the discrepancy suggests that other external factors might have affected the ET-metrics.

Under the assumption that disregarding the principles of proximity and functional grouping did in fact cause a higher cognitive load, the discrepancy could indicate that the ET-metrics can identify aspects of cognitive load that are missed in subjective evaluations. If this would be the case, this highlights a potential strength of ET in detecting differences in cognitive load.

Without this assumption, the discrepancy would indicate that there was

no notable connection between cognitive load and eye-movements. Since the subjective evaluation did not confirm our prediction that the perceived cognitive load would increase with the design changes, it is possible that the significant differences found in the ET-metrics were caused by other influential factors than cognitive load, such as the nature of the task or environmental parameters.

Previous research has shown that the perceived cognitive load was increased when search processes were complicated, similar to what was done in the 'sorting' prototype *(see section 2.5)*. Additionally, there are established connections between the design quality of an interface and cognitive load based on cognitive load theory *(see section 2.3.2)*. With this in mind, it is proposed that the assumption that disregarding principles of proximity and functional grouping causes a higher cognitive load is plausible.

Task performance

During the design of the interface-prototypes, the changes in the designs were kept fairly small as to not lose the aspect of realism. It is suspected that this resulted in differences in cognitive load between the prototypes that were too small to affect the overall task performance *(see section 2.3.3)*. This could explain why no statistically significant differences between any of the three prototypes were seen in the performance metrics. This in turn could offer an explanation to the lack of statistically significant differences in the subjective evaluation, as the prototypes of poor usability were neither more time-consuming nor caused more mistakes than the 'good' prototype and therefore not perceived as more mentally demanding.

Principle of feedback

When comparing the 'good' prototype to the prototype where the principle of feedback was disregarded ('feedback') no statistically significant differences were found based on ET-metrics, performance metrics or the subjective evaluation. The 'feedback' prototype did however cause a high level of confusion in some cases. Some outliers in task performance *(see Appendix A)*, mainly excessive mouse clicks, were found when this prototype was first in the order of tests. The lack of information and previous experience then tricked some users that they had to either double-press on a button or drag-and-drop from the button to the picture to trigger the effect. But when the 'feedback' prototype was second or third in the order, none of the participants had these issues. This indicates that lack of feedback can have an impact on the learnability aspect of usability, although nothing can be said about the connections to eye-tracking metrics based on our findings.

Limitations

During the performance of the experiment with the participants, some issues were noted in the structure of the tasks and tests. In order to ensure consistency, these issues were not resolved so that all participants performed the same tests.

One such issue was that each test was finished with an action in which the participant was instructed to save, share, export or download the postcard. This action can in retrospect be questioned based on its relevance as it had no real connection to the actions performed earlier during the test and that the similarities between some of these icons caused a bit of confusion for some participants.

In the 'sorting' prototype, the total area the buttons were spread out over was somewhat smaller than that of the two other prototypes. This difference could possibly have contributed to the resulting saccade amplitude being lower for this prototype compared to the others.

One of the prompts in the 'good' prototype instructed the participant to change to the picture of the beach from the picture of the skiers after applying a yellow filter to it. This caused some confusion as the filter made the two pictures similar in colour, resulting in some participants thinking they already had chosen the picture of the beach. A similar type of situation with recurring confusion connected to a specific action was not observed in the other two tests and might therefore have affected the performance of the 'good' prototype.

Future directions

The discrepancy between the ET-metrics and the subjective evaluation of cognitive load mentioned earlier is a restricting factor when analysing the results from this thesis. In order to relate the findings from the ET-metrics to cognitive load with greater certainty they would need to align with the results from the subjective evaluation. Based on this, a similar future study where the differences between the prototypes of good and poor usability are exaggerated is encouraged. This could allow for a verification of the influence of design changes on the cognitive load by increasing the chances of finding statistically significant differences in the subjective evaluation. However, exaggerating the differences will negatively affect the aspect of reality which in turn might effect the methods' relevance in a realistic scenario.

The method that was used in this thesis is based on comparison to a known, well-functioning reference. Using ET-metrics to instead evaluate a single design without a reference would require some threshold values. Since

these metrics are dependent on both individual and environmental factors, this approach is deemed highly unlikely. In a commercial setting, the comparative method could instead be used to rank different design suggestions, which could facilitate the design process. Research regarding the demand and applicability of such a comparative method in a commercial setting is therefore encouraged.

In this study, two specific design principles were investigated independently. A possible direction for future research within this field would be to include more design principles, and possibly also combinations of design principles, to examine their possible effect on eye-movements. This could help to further deepen the knowledge of how design influence cognitive processes.

Chapter 6
Conclusion

This master thesis aimed to answer the following problem statements:

- **Can eye-tracking metrics be used as a method to evaluate usability?**
  - Are eye-tracking metrics affected by the design of a user-interface?
  - Can flaws related to specific design principles be identified using eye-tracking metrics?
  - Can eye-tracking metrics be used as an indicator of cognitive load?

Our results indicate that eye-tracking metrics can be affected by the design of a user-interface. There are also indications that specific design flaws can be identified. The data identified a more extensive search behaviour with increased fixation rate, lower saccade amplitude and shorter fixation duration when interacting with the prototype where the design principles of proximity and functional grouping were disregarded. Regarding the principle of feedback, we saw no significant effect in the eye-tracking data distinguishing it from the reference.

Multiple previous studies within this field show results that argue in favour of a connection between eye-tracking metrics and cognitive load. Based on the results of this study, no conclusions can be drawn regarding the connection between eye-tracking metrics and cognitive load. This is because the effect of the designs on the cognitive load was not confirmed by neither the performance based metrics nor the subjective evaluation. Based on this, it is possible to argue for the ability of eye-tracking to capture information about cognitive load missed in subjective evaluation. However, more research is needed to confirm or reject this ability.

Our study shows that this method is able to distinguish extensive search behaviour. Therefore, it could possibly be used to evaluate usability in this particular aspect. However, as usability is a much wider concept, more

research is needed to verify the overall applicability of eye-tracking metrics as a method to objectively evaluate usability.

References

[1] Wang AB. Hawaii missile alert: How one employee 'pushed the wrong button' and caused a wave of panic. The Washington Post [internet]. 2018 Jan 14[cited 2020 Mar 3]. Available from: https://www.washingtonpost.com/news/post-nation/wp/2018/01/14/hawaii-missile-alert-how-one-employee-pushed-the-wrong-button-and-caused-a-wave-of-panic/.

[2] Clarke MA, Schuetzler RM, Windle JR, Pachunka E. Usability and cognitive load in the design of a personal health record. Health Policy and Technology. 2019 Oct 24;9(2020):218-224.

[3] Hollender N, Hoffmann C, Deneke M, Schmitz B. Integrating cognitive load theory and concepts of human-computer interaction. Computers in Human Behavior. 2010 Jun 29;26(6):1278-1288.

[4] Goldberg JH, Kotval XP. Computer interface elavuation using eye movements: methods and constructs. International Journal of Industrial Ergonomics. 1999 Oct;24(6):631-645.

[5] Goldberg JH, Wichansky AM. Chapter 23 - Eye Tracking in Usability Evaluation: A Practitioner's Guide. The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research. Elsevier Science BY: 2003.

[6] Chen S, Epps J, Ruiz N, Chen F. Eye activity as a measure of human mental effort in HCI. Proceedings of the 2011 International Conference on Intelligent User Interfaces, Palo Alto, CA, USA: February 13-16 2011.

[7] Zagermann J, Pfeil U, Reiterer H. Studying Eye Movements as a Basis for Measuring Cognitive Load. CHI EA '18: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada: April 2018. 1-6p.

[8] Widmaier EP, Raff H, Strang KT. Vander's Human Physiology - The Mechanisms of Body Function. 13th Edition. New York, NY: McGraw-Hill; 2014.

[9] Brunyé TT, Drew T, Weaver DL, Elmore JG. A review of eye tracking for understanding and improving diagnostic interpretation. Cogn Res Princ Implic. 2019 Dec;4:7.

[10] Khan MQ, Lee S. Gaze and Eye Tracking: Techniques and Applications in ADAS. Sensors (Basel). 2019 Dec;19(24):5540.

[11] Enderle JD, Bronzino JD. Introduction to Biomedical Engineering. Third Edition. Burlington, MA: Elsevier Inc.;2012.

[12] Carter BT, Luke SG. Best practices in eye tracking research. Int J Psychophysiol. 2020 Sep;155:49-62.

[13] Eto T, Teikari P, Najjar RP, Nishimura Y, Motomura Y, Kuze M, et al. A Purkije image-based system for an assessment of the density and transmittance spectra of the human crystalline lens in vivo. Sci Rep. 2020;10:16445.

[14] Tobii Pro AB (2014). Tobii Pro Lab User Manual (v1.152.1). Tobii Pro AB, Danderyd, Sweden.

[15] Dix A. Human-Computer Interaction. Encyclopedia of Database Systems. Springer, Boston, United States. 2009.

[16] International Organization for Standardization. ISO 9241-11:2018(en). Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts. Geneva: ISO; 2018.

[17] Sharp H, Rogers Y, Preece J. Interaction Design. Fifth Edition. Indianapolis, IN, United States: John Wiley Sons Inc.; 2019.

[18] Lidwell W, Holden K, Butler J. Universal Principles of Design. First Edition. Gloucester MA, United States: Rockport Publishers Inc; 2003.

[19] Norman D. The Design of Everyday Things. 2nd ed. New York, NY, United States: Basic Books; 2013. 10-30 p.

[20] Brody M. User Feedback in User Experience Design. Speckyboy. 2018 Jan 12.

[21] Johnsson J. Designing with the Mind in Mind: Simple Guide to Understanding User Interface Design Guidelines. 3rd ed. Cambridge, MA, United States: Morgan Kaufmann Publishers; 2021. 15-18, 235-243 p.

[22] Szameitat A, Rummel J, Szameitat D, Sterr A. Behavioral and emotional consequences of brief delays in human–computer interaction. International Journal of Human-Computer Studies. 2009 Feb 25; 67(2009) 561-570.

[23] Vaz T. Functional Groups — How can other areas of study help us explain the grouping of elements in design?. Prototypr.io. 2021 Jan 7.

[24] Debue N, van de Leemput C. What does germane load mean? An empirical contribution to the cognitive load theory. Frontiers in Psychology. 2014 Oct 01; 5:1099.

[25] Halarewich D. Reducing Cognitive Overload For A Better User Experience. Smashing Magazine. 2016 Sep 9.

[26] Whitenton K. Minimize Cognitive Load to Maximize Usability. Nielsen Norman Group. 2013 Dec 22.

[27] Chen S. The Construct of Cognitive Load in Interpreting and its Measurement. Perspectives. 2017 Jan 31; 25(4):640-657.

[28] Ramkumar A, Stappers PJ, Niessen WJ, Adebahr S, Schimek-Jasch T, Nestle U, Song Y. Using GOMS and NASA-TLX to Evaluate Human–Computer Interaction Process in Interactive Segmentation. International Journal of Human-Computer Interaction. 2017;33(2):123-13.

[29] Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. Advances in Psychology. 1988;52:139-183.

[30] Zagermann J, Pfeil U, Reiterer H. Measuring Cognitive Load using Eye Tracking Technology in Visual Computing. BELIV '16: Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization. 2016 Oct:78-85.

[31] Hart S. NASA-Task Load Index (NASA-TLX); 20 years later. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2006;50(9):904-908.

[32] Lazar J, Feng JH, Hochheiser H. Research Methods in Human-Computer Interaction. Second Edition. Cambridge, MA, United States: Morgan Kaufmann Publishers; 2017.

Appendix A
Boxplots

A.1 Eye-movement based metrics

A.1.1 Number of fixations

No statistically significant difference was found between the three prototypes, $F(2,58)=1.306$, $p>0.05$ ($p=0.279$) *(see figure A.1)*.

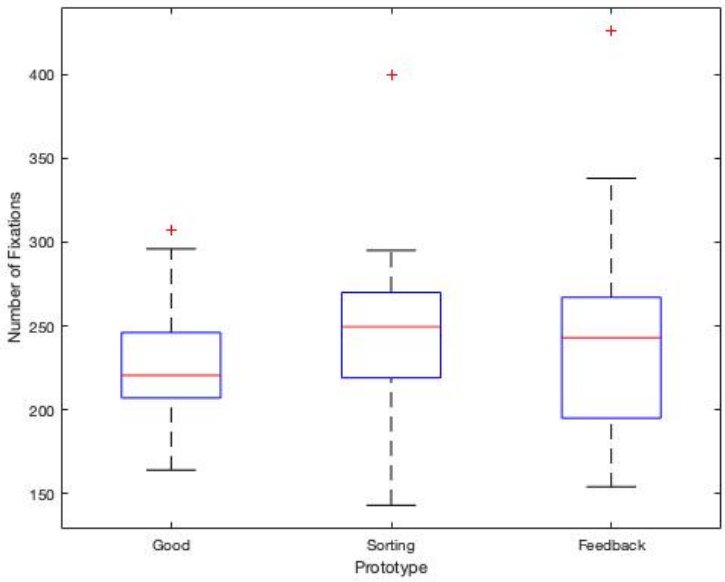


Figure A.1: Boxplot showing the data distribution of total number of fixations with regard to the three prototypes. **Good:** (M=227.93, SD=37.85), **Sorting:** (M=244.90, SD=49.21), **Feedback:** (M=236.43, SD=57.44)

A.1.2 Average fixation duration

Statistically significant difference between the prototypes was found, $F(2,58)=9.297$, $p<0.05$ ($p<0.001$) *(see figure A.2)*. Post hoc-analysis showed that the there was a significant difference when comparing the good prototype to the sorting prototype ($p<0.001$).

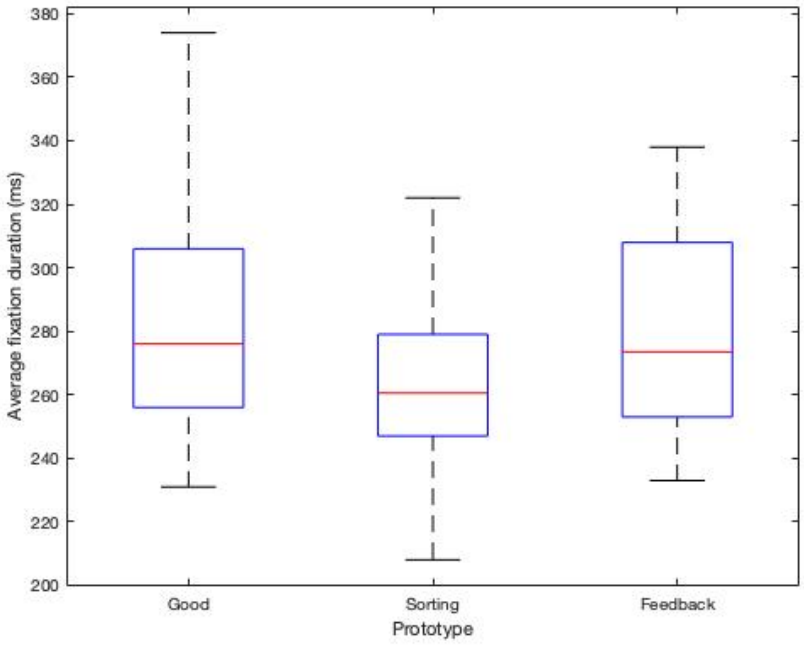


Figure A.2: Boxplot showing the data distribution of average fixation duration with regard to the three prototypes. **Good:** (M=281.77, SD=33.72), **Sorting:** (M=263.77, SD=28.76), **Feedback:** (M=277.57, SD=30.0)

A.1.3 Average saccade amplitude

Statistically significant difference between the prototypes was found, $F(2,58)=12.895$, $p<0.05$ ($p<0.001$) *(see figure A.3)*. Post hoc-analysis showed that the there was a significant difference when comparing the good prototype to the sorting prototype ($p<0.001$).

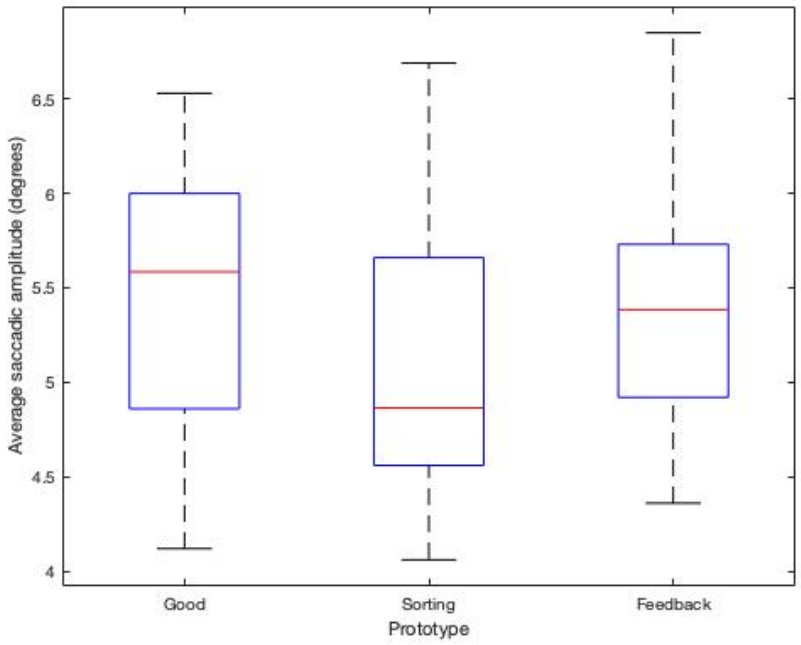


Figure A.3: Boxplot showing the data distribution of average saccadic amplitude with regard to the three prototypes. **Good:** (M=5.49, SD=0.65), **Sorting:** (M=5.07, SD=0.73), **Feedback:** (M=5.44, SD=0.67)

A.1.4 Fixation rate

Statistically significant difference between the prototypes was found, $F(2,58)=11.001$, $p<0.05$ ($p<0.001$) *(see figure A.4)*. Post hoc-analysis showed that the there was a significant difference when comparing the good prototype to the sorting prototype ($p<0.001$).

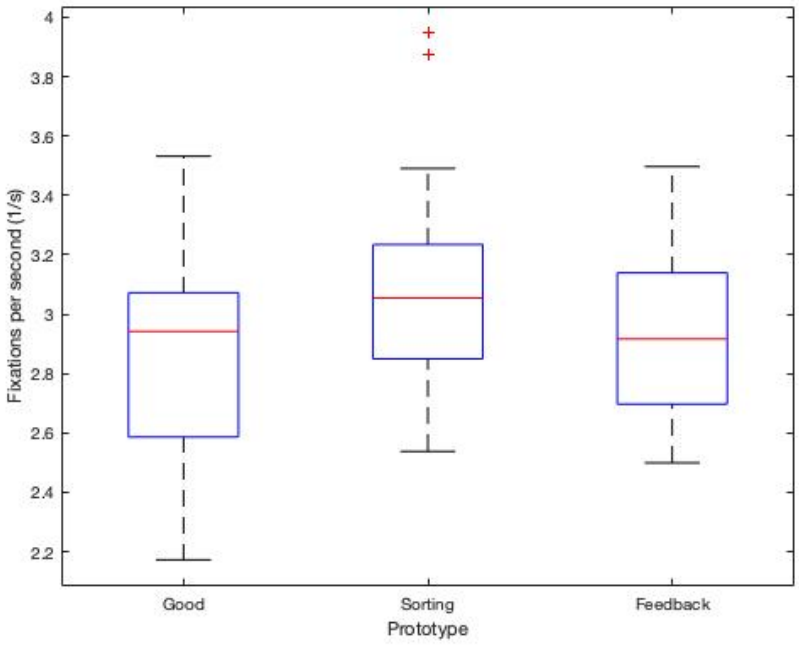


Figure A.4: Boxplot showing the data distribution of fixation rate with regard to the three prototypes. **Good:** (M=2.89, SD=0.33), **Sorting:** (M=3.07, SD=0.34), **Feedback:** (M=2.93, SD=0.3)

A.1.5 Average task amplitude

No statistically significant difference was found between the three prototypes, $F(2,58)=0.453$, $p>0.05$ ($p=0.638$) *(see figure A.5)*.

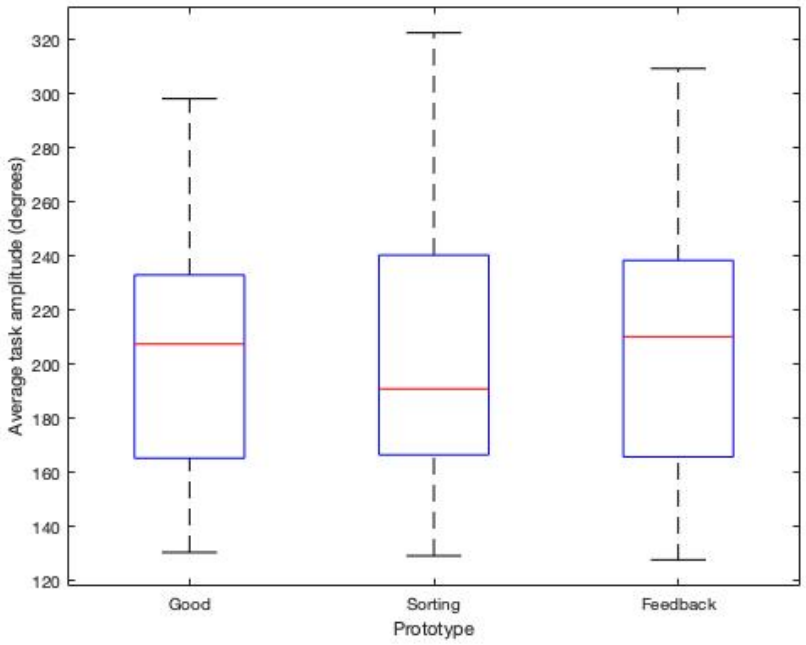


Figure A.5: Boxplot showing the data distribution of average task amplitude with regard to the three prototypes. **Good:** (M=202.81 SD=42.13), **Sorting:** (M=201.52, SD=49.03), **Feedback:** (M=209.6, SD=51.06)

A.2 Performance metrics

A.2.1 Excessive mouseclicks

No statistically significant difference was found between the three prototypes, $F(2,58)=0.733$, $p>0.05$ ($p=0.485$) *(see figure A.6)*.

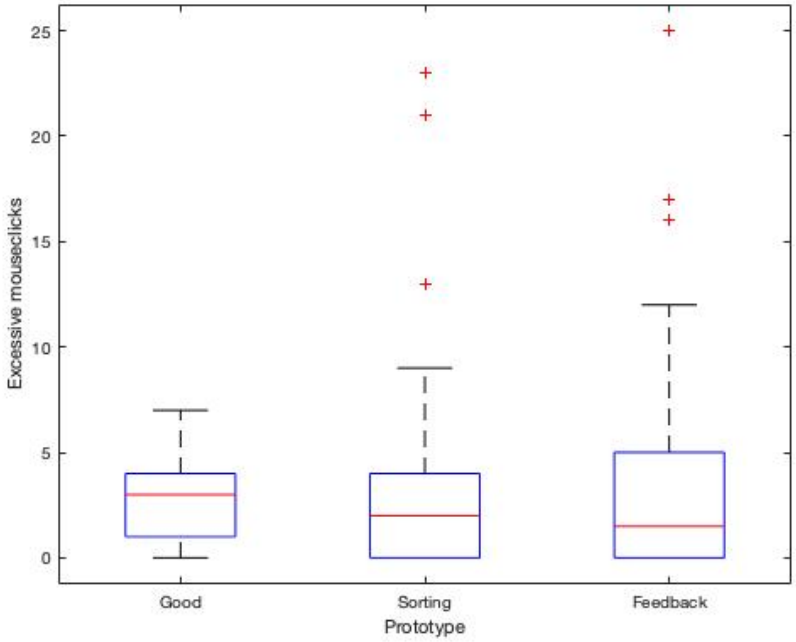


Figure A.6: Boxplot showing the distribution of the number of excessive mouseclicks with regard to the three prototypes. **Good:** (M=2.8, SD=2.06), **Sorting:** (M=3.77, SD=5.81), **Feedback:** (M=4.1, SD=6.05)

A.2.2 Average task duration

No statistically significant difference was found between the three prototypes,$F(2,58)=0.139$, $p>0.05$ ($p=0.870$) *(see figure A.7)*.

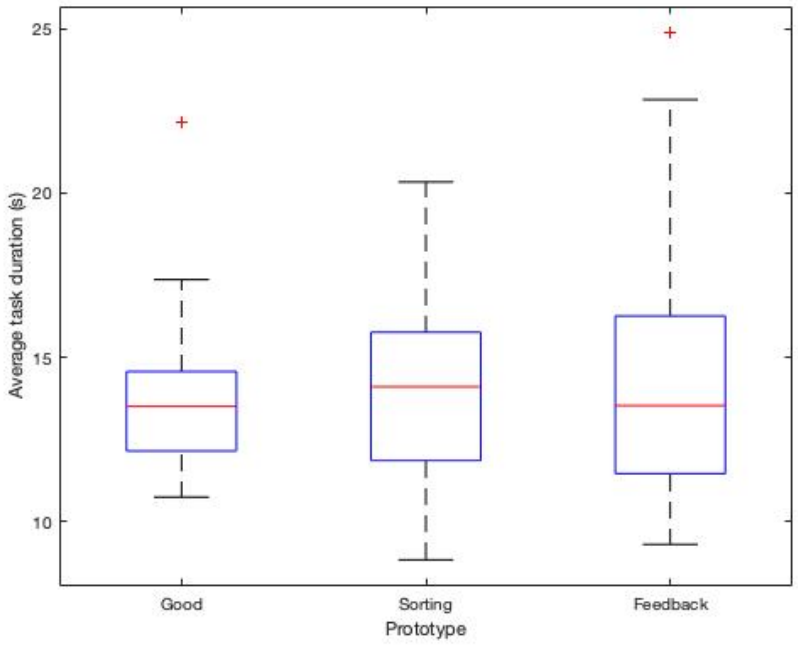


Figure A.7: Boxplot showing the distribution of average task duration with regard to the three prototypes. **Good:** (M=13.75, SD=2.23), **Sorting:** (M=14.02, SD=2.62), **Feedback:** (M=14.06, SD=3.67)

A.3 Subjective evaluation

A.3.1 Mental demand

No statistically significant difference was found between the three prototypes, $F(2,58)=2.284$, $p>0.05$ ($p=0.111$) *(see figure A.8)*.

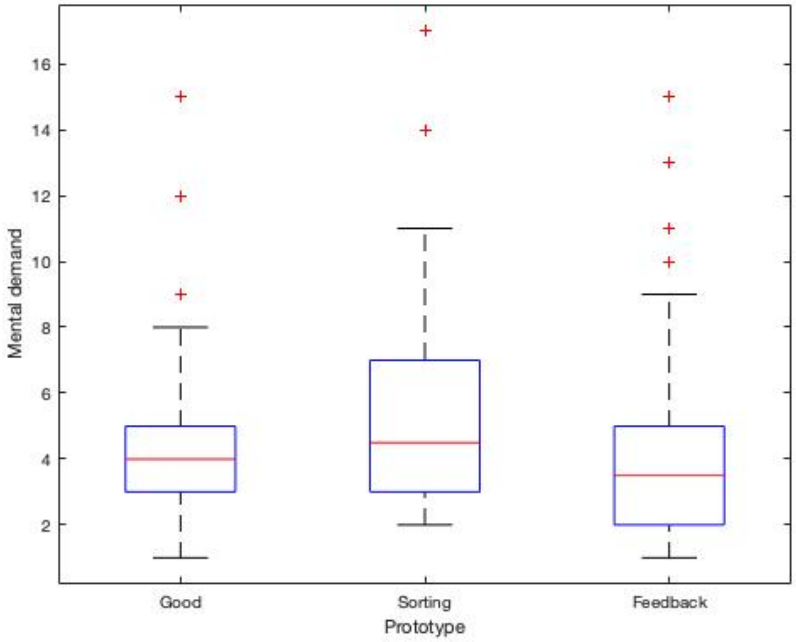


Figure A.8: Boxplot showing the distribution of perceived mental demand with regard to the three prototypes. **Good:** (M=4.63, SD=3.13), **Sorting:** (M=5.53, SD=3.5), **Feedback:** (M=4.67, SD=3.54)

A.3.2 Physical demand

No statistically significant difference was found between the three prototypes, $F(2,58)=0.282$, $p>0.05$ ($p=0.755$) *(see figure A.9)*.

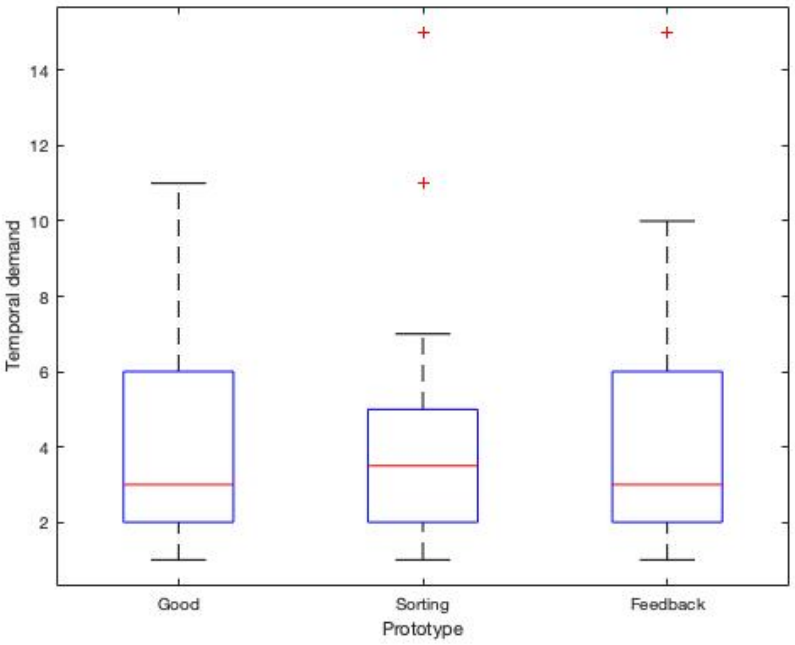


Figure A.9: Boxplot showing the distribution of perceived physical demand with regard to the three prototypes. **Good:** (M=2.03, SD=1.47), **Sorting:** (M=2.07, SD=1.55), **Feedback:** (M=2.17, SD=1.72)

A.3.3 Temporal demand

No statistically significant difference was found between the three prototypes, $F(2,58)=0.162$, $p>0.05$ ($p=0.851$) *(see figure A.10)*.

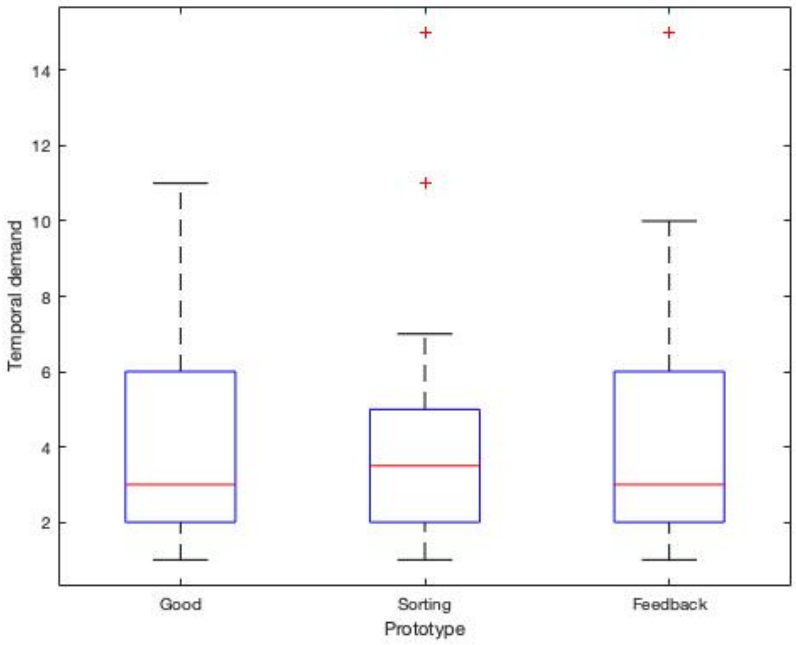


Figure A.10: Boxplot showing the distribution of perceived temporal demand with regard to the three prototypes. **Good:** (M=4.2, SD=2.82), **Sorting:** (M=4.2, SD=3.0), **Feedback:** (M=4.47, SD=3.6)

A.3.4 Performance

No statistically significant difference was found between the three prototypes, $F(2,58)=1.386$, $p>0.05$ ($p=0.258$) *(see figure A.11)*.

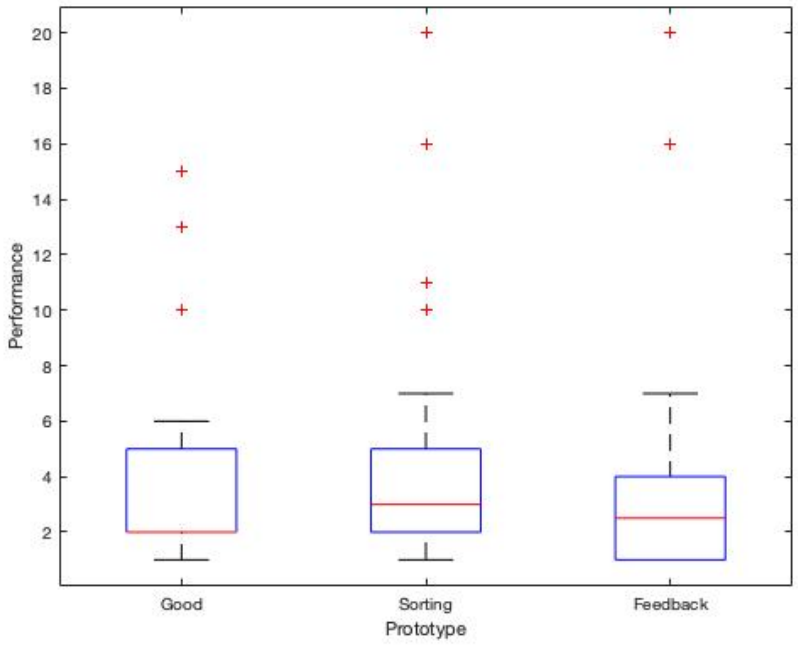


Figure A.11: Boxplot showing the distribution of the participants' perceived performance with regard to the three prototypes. **Good:** (M=3.6, SD=3.46), **Sorting:** (M=4.37, SD=4.51), **Feedback:** (M=3.8, SD=4.25)

A.3.5 Effort

No statistically significant difference was found between the three prototypes, $F(2,58)=0.766$, $p>0.05$ ($p=0.470$) *(see figure A.12)*.

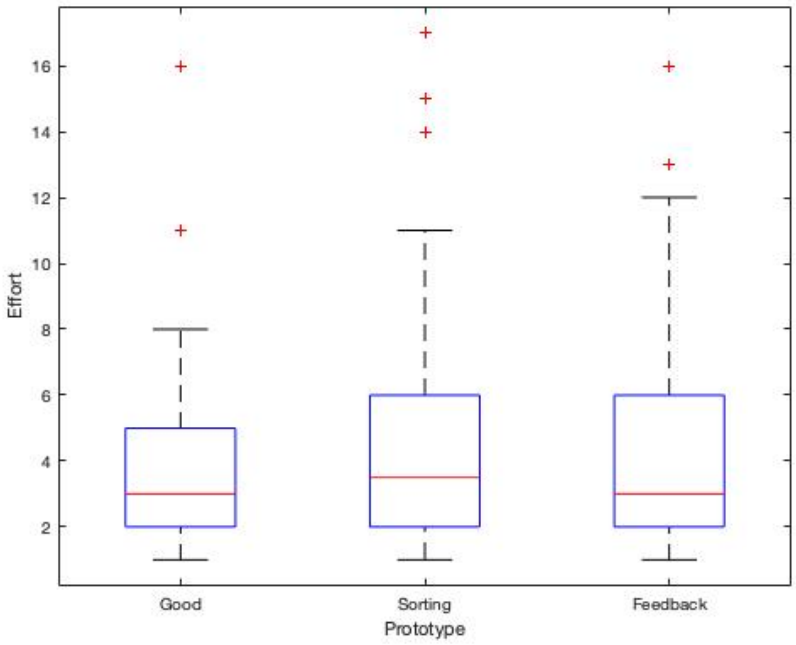


Figure A.12: Boxplot showing the distribution of the participants' perceived effort with regard to the three prototypes. **Good:** (M=4.4, SD=3.43), **Sorting:** (M=4.97, SD=4.44), **Feedback:** (M=4,53, SD=4.34)

A.3.6 Frustration

No statistically significant difference was found between the three prototypes, $F(2,58)=0.555$, $p>0.05$ ($p=0.557$) *(see figure A.13)*.

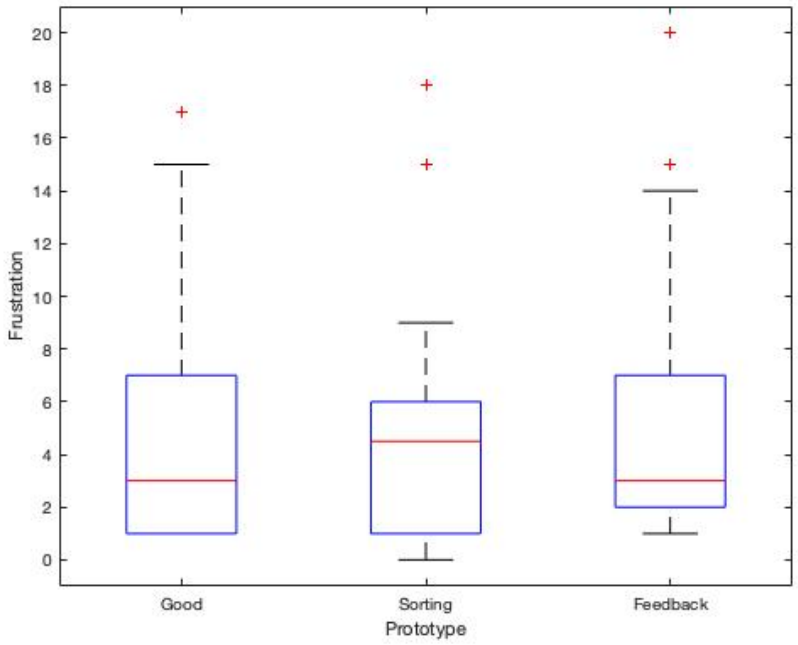


Figure A.13: Boxplot showing the distribution of the participants' perceived frustration with regard to the three prototypes. **Good:** (M=5.13, SD=4.87), **Sorting:** (M=4.6, SD=4.11), **Feedback:** (M=5.33, SD=4.85)

A.3.7 Average score

No statistically significant difference was found between the three prototypes, $F(2,58)=0.348$, $p>0.05$ ($p=0.707$) *(see figure A.14)*.

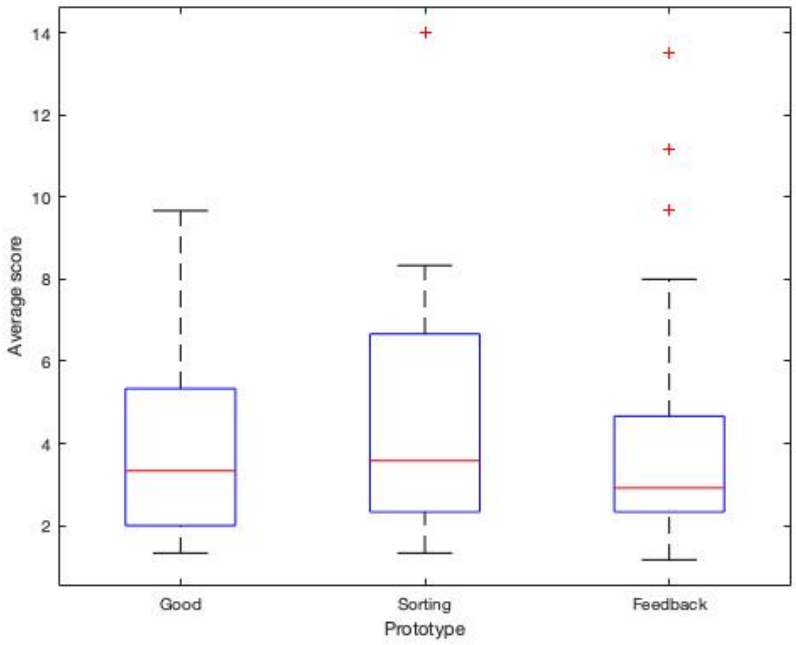


Figure A.14: Boxplot showing the distribution of the average score from the NASA-TLX questionnaire with regard to the three prototypes. **Good:** (M=4.0, SD=2.39), **Sorting:** (M=4.29, SD=2.79), **Feedback:** (M=4.16, SD=3.03)

Appendix B
Consent form

Samtyckesformulär för medverkan under test i samband med examensarbete inom interaktionsdesign

Anton Jigsved & Max Mauritsson

Bakgrund och syfte

Okunskap inom digital arbetsmiljö kan leda till att digitala lösningar utformas på ett bristfälligt sätt, vilket i sin tur kan leda till frustration och problematik och att systemet upplevs som störande snarare än stödjande. Syftet med examensarbetet är att undersöka huruvida eye-tracking kan användas för att utvärdera användbarheten hos digitala användargränssnitt genom att identifiera kognitiva processer. Examensarbetet är tänkt att agera som förstudie inom ett fält med god forskningspotential och eventuellt kommer resultaten från studien publiceras.

Frivillighet

Deltagande i denna studie är frivilligt och Deltagaren kan när som helst avbryta utan att ge någon orsak. Deltagaren har också rätt att begära att redan insamlad information, namn och ögonrörelsedata, om denne raderas.

Resultat

Ögonrörelsedata, ljudinspelning och skriftliga utvärderingar från deltagaren kommer att tas in i forskningssyfte. All data kommer att anonymiseras, förädlas och ligga till grund för studiens resultat. Insamlad data kan komma att publiceras anonymt.

Personuppgifter

I de fall där personuppgifter (namn och kontaktuppgifter) hanteras så kommer dessa anonymiseras, samt i denna förbindelse att behandlas i enlighet med Dataskyddsförordningen (GDPR). Inga personuppgifter kommer överföras till någon tredje part.

Ansvarig

De som ansvarar för hanteringen av datan samt genomförandet av studien är båda M.Sc. studenter inom Medicin och Teknik från Lunds tekniska högskola (LTH). För mer information om projektet eller information om deltagande kontakta: **Max Mauritsson**, *0793-470679*, max_mauritsson@hotmail.se eller **Anton Jigsved**, *0739-587796*, anton.jigsved@gmail.com.

Samtycke

Genom att underteckna denna förbindelse samtycker undertecknad till vad som sägs i denna förbindelse.

Ort och datum: _____

Underskrift: _____

Namnförtydligande: _____

Appendix C

NASA-TLX questionnaire