CHALMERS
UNIVERSITY OF TECHNOLOGY

Eye-Tracking for the Evaluation of Command and Control Interfaces

An Exploratory Study with a Method Proposal for Eye-Tracking Implementation in C2 Usability Evaluation

Master's thesis in Industrial Design Engineering

LYDIA ANTBLAD
SIMON RÅBE ISAKSEN

DEPARTMENT OF INDUSTRIAL AND MATERIALS SCIENCE

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2024
www.chalmers.se

MASTER'S THESIS 2024

Eye-Tracking for the Evaluation of Command and Control Interfaces

An Exploratory Study with a Method Proposal for Eye-Tracking Implementation in C2 Usability Evaluation

LYDIA ANTBLAD
SIMON RÅBE ISAKSEN

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Industrial and Materials Science
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2024

Eye-Tracking for the Evaluation of Command and Control Interfaces

An Exploratory Study with a Method Proposal for Eye-Tracking Implementation in C2 Usability Evaluation

LYDIA ANTBLAD
SIMON RÅBE ISAKSEN

Supervisor: Lars-Ola Bligård
Examiner: Lars-Ola Bligård

Acknowledgments, dedications, and similar personal statements in this thesis, reflect the authors' own views.

Master's Thesis 2024
Department of Industrial and Material Science
Chalmers University of Technology
SE-412 96 Göteborg Sweden
Telephone +46 (0)31-772 1000

Cover:
A collage of a command and control operator with an interface as background. The operator in the image has been generated by an AI image generator.
Cover by Simon Råbe Isaksen, © 2024.

Abstract

Eye-tracking technology has been shown to provide valuable insights into users' behavior and cognitive states, aiding in product and interface design. However, its application to complex user interfaces, particularly command and control systems, remains underexplored. This thesis investigates the use of eye-tracking for evaluating such interfaces, addressing research questions relating to test design, procedure, and data analysis.

An exploratory approach was used, starting with a study of existing literature, followed by the design and execution of user tests in two phases. These tests examined the impact of different design factors and employed various eye-tracking evaluation methods. Data analysis from both qualitative and quantitative perspectives allowed for a comparison of how effectively each method provided insights into the usability of the interface. This culminated in the C2ET (Command and Control Eye-Tracking) method for eye-tracking evaluation of command and control interfaces.

Eye-tracking proved to be a valuable tool for evaluating command and control interfaces, although some limitations exist when using head-mounted eye-tracking devices on these dynamic interfaces. Especially beneficial was the way in which eye-tracking could complement traditional usability methods by confirming, explaining, or expanding upon their insights. The C2ET method offers recommendations for designing, conducting, and analyzing eye-tracking tests, providing designers of these critical systems with a framework on which they can build their own way of working.

The establishment of eye-tracking guidelines in this new field also lowers the threshold for using the technology, thus increasing the accessibility to objective and visual data which can serve as a conduit to communicate the importance of human factors and usability throughout the development process. In addition, this thesis lays a foundation for future eye-tracking research on the topic of evaluating command and control interfaces.

Keywords: command and control, eye-tracking, usability, guidelines, complex user interfaces, evaluation

Acknowledgments

This master's thesis would not have been possible without the support we have received along the way. We would like to direct a special thanks to our examiner and academic supervisor, Lars-Ola Bligård, for his guidance, valuable input and always making time for supervision.

We also express our gratitude to our supervisors at Saab, Erik Halldin and Henrik Nilsson, for your enthusiasm, support, and for being excellent sounding boards throughout this project. Additionally, we would like to thank the Human Factors Integration team and everyone else at Saab Surveillance for making us feel so welcome, with a special thanks to those who contributed to this thesis with their participation in our user tests.

Thank you, Saab, for the opportunity to conduct our master's thesis at the company, providing us with knowledge and experiences that we will carry with us.

Lastly, we extend our heartfelt thanks to Chalmers University of Technology for an incredible five years, equipping us with the skills and knowledge that will guide us as we embark on the next chapter of our lives.

Lydia Antblad *Simon R. Isaksen*

Table of Contents

01

Introduction

Eye-tracking technology has been around for decades, during which researchers have sought to use it to study the usability of products and interfaces. Although eye-tracking has been hailed as a powerful tool, allowing insights into users' behavior and cognitive state, its use has not been as extensive as might be expected. High costs, complex procedures, and data that is difficult to analyze have created a high threshold that has hindered its widespread implementation in the industry, outside academic research settings. With eye-tracking technology becoming more accessible, it is now possible for companies to employ it to a higher extent. By integrating eye-tracking into the design process, organizations can better understand how users interact with their products, uncovering valuable insights to inform decision-making.

This thesis is a collaboration with Saab AB, a company that works actively to achieve the best human-machine interaction possible and is continuously looking for new ways to improve the evaluation of its products. Using eye-tracking for the evaluation of interfaces is a relatively new undertaking at the business unit Saab Surveillance. Because of this, a need for a clear framework on when and how the technology can be used has become apparent.

Despite the large collective research that has been done on eye-tracking and its uses, the extent to which this research has been done on complex user interfaces is limited. Specifically in command and control, where a large part of Saab Surveillance's product portfolio lies, research does not appear to have shown how eye-tracking can be applied. It is therefore interesting, both from a company and research perspective, to study how these interfaces can be evaluated using eye-tracking.

1.1. Aim

This thesis aims to research the best practice use of eye-tracking in the evaluation of command and control interfaces and determine how to interpret its data to draw conclusions about usability. Insights will be used to formulate a context-based method for the use of eye-tracking in the development and evaluation of command and control interfaces at Saab Surveillance.

1.2. Goal

The primary goal of this thesis is to supply the Human Factors Integration team at Saab Surveillance with practical and useful guidelines for how to utilize eye-tracking in their work with evaluating and developing the usability of human-machine interfaces.

1.3. Research Questions

To aid in the completion of the goal, this study has three research questions:

RQ1: How should user tests using eye-tracking be designed to evaluate the usability of command and control interfaces effectively?

RQ2: How should user tests on command and control interfaces using eye-tracking be performed to attain useful results?

RQ3: Which data and methods are most useful for efficiently evaluating the usability of command and control interfaces using eye-tracking?

1.4. Deliverables

The deliverable for this master's thesis is a method for understanding when and how to use eye-tracking in the evaluation of command and control interfaces, consisting of guidelines for how to:

- plan and set up eye-tracking tests;
- perform eye-tracking tests and which procedures to follow; and
- interpret results from different evaluation methods.

1.5. Research Contribution

This master's thesis contributes to research in the field of eye-tracking by addressing the knowledge gap pertaining to complex user interfaces, specifically command and control interfaces, with dynamic elements constantly changing and moving. Unlike the controlled settings of typical lab environments where much previous eye-tracking research has been conducted, this study delves into the complexities inherent in real-world product development settings. The contribution thereby also lay in the possibility of extending beyond academic research, to encompass practical application and insights into the feasibility of applying theoretical knowledge to the testing and refinement of real products.

1.6. Industry Contribution

The usability of command and control interfaces, which are generally complex and highly technical, has been known to be problematic (McNamara et al., 2019). By contributing insight into how eye-tracking can be used to evaluate these interfaces, this thesis will offer designers a new tool to use for continuous improvement.

Saab Surveillance has identified the need for a framework for optimal use of eye-tracking in their context. With a clear procedure for its employment, the Human Factors Integration team will have better access to a technology that enables them to support their findings with objective and visual data. This will further the strive to keep human-machine interaction and usability a central part of the development process and facilitate cross-functional cooperation.

1.7. Delimitations

To confine the scope of the study, it had the following delimitation:

- The only eye-tracking device used was Tobii Pro Glasses 3 and its associated software.
- The study was exploratory in nature, trying to explore possible applications of the technology rather than trying to prove anything definitively.
- Findings from previous research were not questioned nor confirmed, the study instead tried to implement already gained knowledge practically.
- The developed method was not tested nor implemented after its completion.
- Testing was limited to a single command and control interface.

1.8. Outline of Thesis

The following is the content of the chapters in this thesis:

Chapter 1: Introduction – Introduces the topic of the thesis as well as the aim, goal, research questions, deliverables, contribution and, delimitations of the study.

Chapter 2: Background – Describes the context of the study including the products and technology involved as well as the research approach that was applied.

Chapter 3: Study of Literature – Presents previous research and theory related to eye-tracking and usability.

Chapter 4: General Process & Execution – Describes the process and execution of the study in general, including the method development process.

Chapter 5: Testing Phase One – Presents the first testing phase of the study, its execution, and the findings from the user tests A and B. The chapter concludes with the learnings that were transferred to phase two.

Chapter 6: Testing Phase Two – Presents the second testing phase of the study, its execution, and findings from user tests C and D. The subchapter for each test concludes with a comparison between the contribution of the different evaluation methods employed in the tests respectively.

Chapter 7: Review of Test Design & Evaluation Methods – Discusses the study's test design and the results of the different evaluation methods employed in both testing phases.

Chapter 8: The C2ET Method – Presents the developed method. Includes a description of aspects to consider when designing and performing eye-tracking tests as well as how to interpret the data.

Chapter 9: Discussion – Discusses the results, the developed method as well as the execution of the study. Whether research questions have been answered is also discussed. Lastly, ethical considerations are presented, and opportunities for future work.

Chapter 10: Conclusion – Describes the general conclusions of this thesis and future opportunities.

02

Background

This chapter aims to provide an understanding of the context of this study, specifically Saab Airborne Surveillance, along with a description of the products and technology involved. Finally, it offers a description of exploratory research to establish the basis for the study's approach.

2.1. Saab Airborne Surveillance

Saab AB is a Swedish aerospace and defense company producing technological solutions in military defense, aviation, and civil security. Saab Surveillance, a business unit within Saab, develops ground and air-based sensor systems. Included in the product portfolio of Saab Surveillance is the GlobalEye Airborne Early Warning and Control (AEW&C) solution (see Figure 1).

The Human Factors Integration (HFI) team at Saab Surveillance works with, among other things, evaluating and improving the products from a human-machine interaction perspective. The HFI team has expressed a need to explore how the utilization of eye-tracking could be optimized for their current context of command and control interface development.

Figure 1

GlobalEye AEW&C in a coastal environment

Note. (Saab AB, 2019)(Photo: Saab AB, Copyright Saab AB)

2.1.1. System Overview

AEW&C systems use command and control (C2) interfaces to connect operators to its suite of sensors. In this study, a C2 application was used to conduct tests and evaluations. This application should be considered a complex system because of the vast amount of diverse information it handles and its elaborate operational procedures. Given its complexity, the C2 system is commonly used by several highly trained and experienced operators simultaneously, to fully utilize its functionality. Tasked with continuously processing information from the C2 system, the operators need to rapidly assess situations, orient themselves, make decisions, and act. This is a process closely tied to the OODA (Observe, Orient, Decide, Act) loop framework (Osinga, 2006). Workload and operational intensity vary, however, with operators performing everything from peacetime surveillance to wartime fighter control.

2.2. Command and Control Interfaces

Command and control is a concept that had its beginnings in the early days of warfare (Alberts & Hayes, 2006). While the purpose of C2, to combine the efforts of various entities to achieve a specific task or goal, has largely remained consistent throughout history, how its functions have been accomplished has developed significantly.

Today, digital interfaces exist to assist in C2. These C2 interfaces can for example include functionality such as route planning, sensor control, target detection, and target identification (Saab, n.d.). The contexts in which C2 interfaces are used, be it intelligence, surveillance, or reconnaissance, can be mission-critical and high-pressure for operators (McNamara et al., 2019). Effective user interfaces that allow for efficient ways of working in these settings are therefore vital. C2 interfaces usually differ from traditional desktop applications by being information-dense, highly technical, and filled with multiple windows with settings. This means that these interfaces are generally complex systems with a steep learning curve, requiring experienced operators. A complex system is characterized

by Snyder et al. (2011, p. 468) as one that "is composed of interconnected parts that as a whole exhibit one or more properties not present in the individual parts alone".

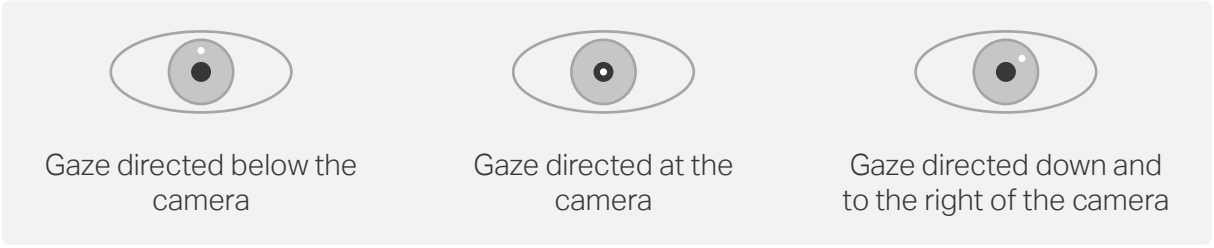
The challenge of designing C2 interfaces with good usability has been stated previously by McNamara et al. (2019), adding that even experienced crews can have trouble understanding these complex systems. Overly complex systems can cause operators to be exposed to a high cognitive workload, and as a result, need to divert cognitive resources away from tasks to deal with the inadequacy of their interface. However, McNamara et al. (2019) also state that these difficulties are not necessarily intrinsic to these types of interfaces. If efforts are made to follow conventions in usability and human-machine interaction, there is no reason many of these issues could not be fixed.

2.3. Eye-Tracking Technology

Eye-tracking has been around for over a hundred years and several techniques have been developed to measure the movement of the eye either by measuring the point of regard (where one is looking) or the motion of the eye relative to the head (Duchowski, 2017). Eye movement measurement methodologies can be categorized into four broad categories: Electro-OculoGraphy (EOG), scleral contact lens/search coil, Photo-OculoGraphy (POG) or Video-OculoGraphy (VOG), and lastly video-based combined pupil and corneal reflection. Today, the most widely used technique is within the latter category: pupil center corneal reflection (PCCR). The corneal reflection, also known as the 1st Purkinje reflection or glint, is measured relative to the location of the pupil center (see Figure 2). To distinguish the reflection from surrounding light, infrared light is used and placed at a fixed position relative to the eye. The reflection from the infrared illuminators along with the pupil's position is used to calculate the point of regard.

Figure 2

Corneal Reflection Position According to Point-of-Regard



The earliest eye trackers were built in the late 1800s by Edmund Huey, however, they were difficult to construct and required invasive methods involving direct contact with the cornea (Holmqvist et al., 2011). The first non-invasive eye measurements using corneal reflection were performed in 1901 by Dodge and Cline. In the 1950s the technique for recording eye movements from the reflection of an external light source was refined and developed further using contact lenses with various mechanical or optical devices to gain more sensitive measurements (Robinson, 1968). It was also in the 1950s that the first usability evaluation using eye-tracking was performed by Paul Fitts and his colleagues, a study with impact still today (Fitts et al., 1950). They studied the pilot's eyes as they used cockpit controls and instruments to land an airplane. In the 1970s great technological advancements in eye-tracking were made for more precise eye monitoring,

enabling higher accuracy of data and fostering more research on what the human eye can reveal about human perceptual and cognitive processes (Jacob & Karn, 2003). However, similar advancements were lacking with regard to data processing, resulting in a decreasing rate of research on the use of eye-tracking. Today, with the advancement of remote and wearable eye trackers along with image processing, sensitive measurements with greater credibility can be obtained in a more accessible and less invasive manner (Duchowski, 2017). Besides the advancement of eye-tracking technology, recent software development has assisted the interpretation of data and made research using eye-tracking more feasible.

The first eye trackers in the 19th century were used to analyze the process of reading (Pluzyczka, 2018). Today, eye-tracking is used in a variety of research fields including psychology (Fernandez-Lanvin et al., 2023), marketing (Xie et al., 2023), human-computer interaction (Li et al., 2022), and user experience (Pirus et al., 2023). There exist various types of eye-tracking devices suitable for different application domains:

- head-stabilized trackers that require the head to be still during data collection;
- remote trackers that do not require contact with the participant, for example, screen-based trackers;
- mobile eye trackers that are head-mounted and allow for in-action tracking, like wearable eye-tracking glasses; and
- integrated systems, for example, virtual reality devices (Mento, 2020).

Tobii is a leading company in eye-tracking and has for the last twenty years developed eye-tracking technology for numerous contexts such as consumer research, user experience, healthcare, gaming, and the automotive industry (Tobii, n.d. -a). They offer three different types of screen-based eye trackers and one wearable eye tracker along with various accessories. Besides these products, they offer Tobii Pro Lab, a software application for research and analysis. This project will be conducted using Tobii Pro Glasses 3 and the mentioned software.

The technology behind Tobii Pro Glasses 3 is based on video-based pupil and corneal reflection eye-tracking (Tobii, 2023c). The glasses have 4 cameras (2 per eye) and 16 illuminators (8 per eye) integrated into the lenses (Tobii, 2024a).

The system components of Tobii Pro Glasses 3 include the head unit comprised of the wearable eye-tracking glasses and the recording unit connected to the head unit via a cable (Tobii, n.d. -b). The recording unit holds a battery and an SD card. Connected to the recording unit is the computer via an ethernet cable. The Tobii Pro Glasses 3 controller application on the computer manages the glasses and allows you to calibrate the glasses, start and stop recording, prepare events ahead of recording, place events during a recording, and view old recordings. Tobii Pro Lab is a secondary application and is used after recording to analyze the results of the eye-tracking recording. In addition to the glasses, there are lenses of various strengths that can be inserted. Nose pads can also be attached to ensure the glasses fit comfortably.

Tobii Pro Lab allows for three different types of analysis: metrics export, metric visualizations, and visualizations (Tobii, 2024b). Metrics export allows any chosen raw data to be exported to a spreadsheet format. Metric visualizations, on the other hand, is Tobii Pro Lab's internal analysis tool that can create graphs and visualize data points quickly. Visualizations enable the creation of heat maps, scan paths, and bee swarms, for image-based evaluation. Data can either be separated by participant and recording or aggregated based on participant characteristics.

Times of interest (TOIs) are a concept inside Tobii Pro Lab that allows the user to specify intervals in which data are to be visualized or exported (Tobii, 2024b). This can be used to isolate specific moments in a longer recording, or for splitting a recording up into tasks. To create a TOI, one or two event markers must be added, around which the TOI is defined.

Assisted mapping is a tool in Tobii Pro Lab that automatically tracks the head-mounted recording from glasses onto a still image (Tobii, 2022). This can either be onto a screenshot from the recording or a picture taken separately. Mapping can also be done manually by marking the location of the gaze for each frame. Mapping is necessary for image-based visualizations but can also be used for the creation of areas of interest (AOIs). AOIs can be created statically onto a still image, or dynamically, following the video. The dynamic AOIs require the practitioner to create keyframes so the shape follows the movement of the video.

2.4. Exploratory Research

Exploratory research intends to explore research questions and does not entail final answers or conclusive solutions to problems (Casula et al., 2021). The research is often conducted by studying a problem that has not been clearly defined to understand the nature of the problem. It is a method often employed to tackle new problems on which little or no previous research has been done and is effective in laying a foundation on which future studies can build. The downside of exploratory research is the lack of conclusive results that it generates and the difficulty of summarizing the qualitative data in an objective manner (Yeaton et al., 1995). Moreover, exploratory research commonly uses small sample sizes, which increases the risk of the result being non-representative on a larger scale.

03

Study of Literature

This chapter covers various research areas relevant to eye-tracking and usability research including eye physiology, the eye-mind hypothesis, eye-tracking methodology, interpretation of eye-tracking data, eye-tracking in design, usability, user tests, and think-aloud protocol. The theories presented in this chapter will lay the groundwork for the research in this thesis.

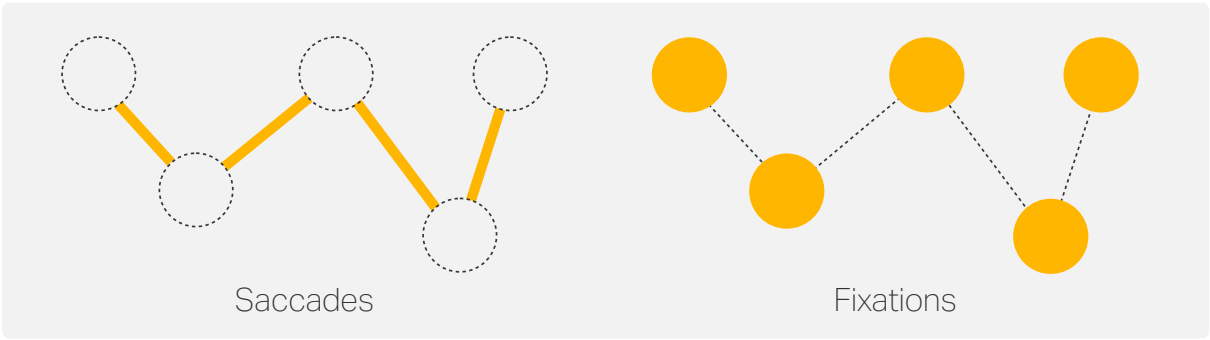
3.1. Eye Physiology

The human vision system is a complex one and consists of both physiological and neurological processes (Holmqvist et al., 2011). For the purpose of understanding eye-tracking, some of the physiology of the eye and its basic movements need to be understood. The human eye lets light in through the pupils, which changes size depending on the amount of light. As the amount of light decreases, the pupil gets larger. Photoreceptors at the rear interior surface of the eye, in the retina, convert light energy to electrical impulses also known as neural signals. These photoreceptors are classified into two categories: rods and cones (American Academy of Ophthalmology, 2018). Rods are sensitive to light and support vision under dim light conditions. They also allow us to distinguish large shapes and motion and support our peripheral vision. Cones respond to chromatic light and support our color and daylight vision. At the back of the eye, there is a small area called the fovea that is densely packed with cones allowing us to focus on objects and their details (Holmqvist et al., 2011). In other words, the fovea allows for high visual acuity and high resolution. This focus area is about the size of one or two words on a computer screen (Nielsen & Pernice, 2010). For example, when reading one must foveate i.e., move one's eyes so that the light from the word falls directly on the fovea, to be able to read the word.

The eye can move in three dimensions (Holmqvist et al., 2011). However, the eye does not pan smoothly to observe but in spurts. Rapid movements for repositioning the fovea to a new location are called saccades (see Figure 3). These can be either voluntary or reflexive. The duration of saccades typically ranges from 30 – 80 ms during which time we are effectively blind a phenomenon called saccadic suppression (Nielsen & Pernice, 2010; Holmqvist et al., 2011). The state in between saccades, when the eye stops and focuses on an object, is called a fixation (see Figure 3). This is when visual information is acquired and lasts from milliseconds up to several seconds, but typically 200 – 300 ms. Fixations do not mean the eye is completely still, in fact, it is characterized by micro-movements in the form of tremors, drift, and microsaccades (Riggs & Ratliff, 1951). A tremor is a small movement that helps retain visual acuity during fixations. Drifts are slow, irregular movements of the eye from the center of fixation and microsaccades are quick movements that bring the eye back to its original position (Holmqvist et al., 2011). Smooth pursuit is another type of eye movement that the eye performs when following an object, e.g., a bird in the sky.

Figure 3

Visual Representation of Fixations and Saccades



Human visual acuity is a visual performance parameter that refers to the clarity of vision and one's ability to recognize details with precision (Nielsen & Pernice, 2010). The visual acuity declines towards the periphery and since eye-tracking records what the user's foveal vision fixates on it is not possible to determine what the user perceives from the periphery. Nonetheless, peripheral vision plays a central role in motion detection (Duchowski, 2017). And even though peripheral vision does not enable perception of details or for example, reading, it is sufficient for determining general shapes and colors of elements (Nielsen & Pernice, 2010).

3.2. The Eye-Mind Hypothesis

The eye-mind hypothesis is the foundation on which eye-tracking data is interpreted in usability studies (Nielsen & Pernice, 2010) and implies that people are thinking about what their eyes are fixated on. The hypothesis relates cognitive processes with the direction of gaze i.e., attention with fixation (Just & Carpenter, 1980). It is an assumption with limitations as it is possible to imagine an object without having to look at it (Nielsen & Pernice, 2010). A person's attention is not always on whatever their foveal vision is fixated on. Nonetheless, the hypothesis holds true often enough for eye-tracking to be a useful method to tell us what users pay attention to.

Eye-tracking data can show what a user is looking at but is insufficient in answering questions about what the user is thinking about when an element is being processed, why they pay their attention to that element, or how they interpret the element. It does not convey full information about the cognitive process behind fixations. For example, fixation data can indicate the following processes:

- difficulties in extracting information (Jacob & Karn, 2003)
- heightened cognitive attention (Andrá et al., 2013)
- mental calculation (Hartmann et al., 2015)
- bored staring (Schindler & Lilienthal, 2019)

Models of cognitive processes cannot solely rely on eye-tracking data as the relationship between individual eye movement and specific mental operations cannot be unambiguously determined (Epelboim & Suppes, 2001; Nielsen & Pernice, 2010). Therefore, interpreting data points on fixations which indicate attention, and deciding on the implications of the data should be done in conjunction with other knowledge, as attention does not always imply understanding. Consequently, it is not feasible to determine exclusively from eye-tracking data whether a user comprehends an element as intended by the designer. In summary, while eye-tracking can identify which parts of a stimulus a person is processing, it cannot reveal the types of processing that occur (Cullipher et al., 2018).

3.3. Eye-Tracking Methodology

Below is a description of what previous research has highlighted as particularly important to consider when user tests with eye-tracking are being performed.

3.3.1. What Types of Questions Can Eye-Tracking Answer?

Eye-tracking, when used as a quantitative method can, according to Cullipher et al. (2018), directly answer the following questions:

- "At what part of the stimulus is the participant looking?"
- "How much time does the participant spend looking at a particular part of the stimulus?"
- "Does the participant look at a particular part of the stimulus more than the others?"
- "In what order does the participant view the various components of the stimulus?" (Cullipher et al., 2018, p. 2)

Following the eye-mind hypothesis, eye-tracking can also provide some information about the underlying cognitive processes of a participant interacting with visual stimuli (Nielsen & Pernice, 2010), hence, according to Cullipher et al. (2018), answer questions such as:

- "What part of the stimulus does the participant spend most of their time processing or trying to interpret?"
- "How much time does the participant spend processing different parts of the stimulus?"
- "In what order does the participant process information presented in the stimulus?" (Cullipher et al., 2018, p. 3)

However, due to the limitations of eye-tracking data to provide information about what types of cognitive processing that occur, additional evaluation methods are necessary (Cullipher et al., 2018). Methodological triangulation, as further described in Chapter 3.3.4, contributes to better interpretation and understanding of the quantitative findings acquired with eye-tracking.

3.3.2. Test Design

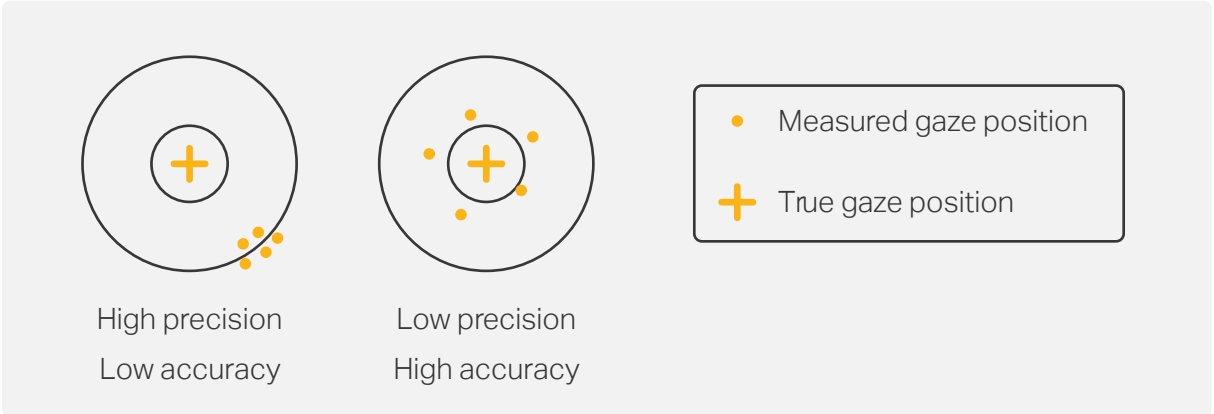
There are several factors that need to be considered when designing eye-tracking tests. Some of the most common factors to consider are presented below.

Accuracy of Measurements

The quality of eye-tracking data is a combination of accuracy and precision (Holmqvist et al., 2011). Accuracy meaning the difference between true gaze position and recorded gaze position, and precision meaning the eye tracker's ability to reproduce a measurement (see Figure 4). Eye-tracking data of high quality naturally have both high accuracy and precision.

Figure 4

Illustration Showing the Difference Between Precision and Accuracy



The gaze sample percentage is an indicator of the quality of the eye-tracking measurement (Havanki & Hansen, 2018). When the tracker cannot detect a participant's pupil it interprets this as invalid or missing data.

Participants

Data collection from a group of participants can be done either by employing a between-subjects or a within-subjects test design. Since eye-tracking measurements are individual, it is advised that eye-tracking studies use a within-subject design to enable comparison between individuals interaction with different test objects (Duchowski, 2017).

Nielsen and Pernice (2010) list the following recommendations for the number of participants in their guide to eye-tracking research:

- Eye-tracking aimed at generating heatmaps and these are the main deliverable: 30 participants.
- Watching gaze replays: Six participants.

The following participant criteria should be met to ensure the quality of eye-tracking recordings (Tobii, 2023b):

- Inform participants to avoid wearing heavy eye makeup or big eyelashes.
- Avoid participants whose pupils are constantly dilated (e.g., because of medication).
- Exclude participants who have had eye surgery.
- Exclude participants with cataracts, amblyopia, strabismus, nystagmus, and eyelid ptosis.

Environment

The following environmental considerations should be made to ensure the quality of the Tobii Pro Glasses 3 eye-tracking (Tobii, 2023b):

- Ideal illuminance should be around 300 lux.
- Avoid sources that produce flickering light.
- Avoid sunlight and devices that emit near-IR light (e.g., halogen lamps) can disrupt eye-tracking.

Task Selection

Eye movements are task-dependent so the nature of the task given to the test participant will influence the outcome of the eye-tracking (Duchowski, 2017). Our gaze is directed by a combination of bottom-up (stimulus-driven) and top-down (goal-driven) cognitive processes. Thus, tasks need to be carefully selected so that cognitive processes can be attributed appropriately.

Stimuli

Deciding on stimuli is a vital point in the design of an eye-tracking study (Havanki & Hansen, 2018). It must be made in a way that allows for the task to be completed without distraction or influence from other factors to allow for data analysis and visualization. Stimulus is, according to Havanki and Hansen (2018, p. 26), "an image or event that elicits a behavioral or emotional response from a participant" with the main purpose of answering the research questions. Within the stimuli, are elements that are of interest to the research as well as distractors with features similar to those of the elements. Moreover, there is the background that includes all remaining objects of the stimulus.

According to Havanki and Hansen (2018), stimuli can be categorized into eight different categories depending on three factors:

- Type of visual image, static or dynamic:
  - Static means still images are projected on a screen that do not change during viewing.
  - Dynamic means moving images, which could be real-world environments through the use of wearable eye trackers, or virtual, augmented, or mixed reality.
- Level of interaction, passively viewed or interactive:
  - Passive means only viewed by participants, not allowing for any manipulation.
  - Interactive means for example simulations, websites, or video games where the participants can change what they are viewing through their actions. A test where the participant is given more freedom in the interaction with the stimuli can make comparisons between tests difficult.
- Display properties, 2D or 3D:
  - 2D means for example screens or pictures.
  - 3D means real-world environments where the eye tracker considers depth.

The C2 simulator used in this study can be categorized as dynamic, interactive, and two-dimensional (DI2).

Laboratory Versus Field Research

Eye-tracking often has pragmatic constraints making it more suitable to perform research in a laboratory setting than "in the field" (Duchowski, 2017). Computers, cables, and other equipment as well as the less controllable nature of real-world tests make it at times impossible and other times unsuitable and inefficient if the goal is to obtain reliable data. However, from a usability point of view, ecological validity holds significant value in fulfilling the aim of understanding the human-machine interaction in its real context. With the arrival of portable eye trackers, it is now feasible to perform measurements outside of the laboratory setting and achieve ecological validity at a level that has previously been unattainable. A study investigating how movement influences the accuracy of the Tobii Pro Glasses 3 showed no significant effect on the accuracy of the eye-tracking

measurements (Onkhar et al., 2023). The technology therefore has the potential of providing insight into user interaction in situations when other usability methods may be inapplicable. Nonetheless, eye-tracking in field studies, for example in aviation settings, entails challenges that laboratory settings do not, as further described by Pignoni and Komandur (2022).

3.3.3. Data Analysis

When relating eye-tracking data to cognitive theory one can interpret fixations from two perspectives, according to Jacob and Karn (2003):

- Top-down based on cognitive theory or design hypothesis; and
- Bottom-up based on observations of data without predefined theories.

A top-down approach requires either a hypothesis or a data-driven search for fixation patterns. Top-down, based on cognitive theory means that, for example, longer fixations on an element indicate difficulty interpreting the element (Jacob & Karn, 2003). Furthermore, a top-down interpretation based on a design hypothesis means that the data is analyzed to see if the hypothesis holds true or not. For example, people will detect a notification faster if it is placed in the center compared to at the bottom of the page. A bottom-up is of a more exploratory nature. For example, the participant is taking much longer than anticipated to find this element. We wonder where they are searching.

Finally, eye-tracking generates large amounts of data, so it is essential to find ways of analyzing it automatically, not only to save time but also to minimize the chance of errors that manual data processing may bring (Poole & Ball, 2006).

3.3.4. Method Triangulation

Method triangulation is the application of more than one data source and data collection method in the study of a phenomenon (Patel & Davidson, 2019). The outcome may differ or coincide between the methods, either way, the result is of interest, and together the information gives a fuller picture of the research topic. If a finding is based on only one data source, it may suffer from limitations connected to the method itself or the application of it. In the case of eye-tracking data, auxiliary data serves as a way to deepen and to cross-validate it. Holmquist et al. (2011) lists several complementary methods to eye-tracking:

- interview
- think-aloud
- objective usability measurements (for example number of errors and task completion time)
- questionnaires (for example the Likert scale)
- biometric measurements (for example galvanic skin response (GSR) and electroencephalography (EEG)) (Holmqvist et al., 2011)

When triangulating eye-tracking results with other data collection methods one must consider how these may impact the eye-tracking data and how the auxiliary data will be related to the eye-tracking data (Cullipher et al., 2018). For example, concurrent verbalizations have been shown to cause reactivity by either improving or impeding the task performance, impacting the eye-tracking data (Russo et al., 1989).

Nonetheless, verbal data, in particular, has the potential to disambiguate eye-tracking data as it provides insight into the mental processes that participants experience while performing a task or inspecting a stimulus (Holmqvist et al., 2011). The question of whether verbal data adds value to eye-tracking data has been a topic of research and is further discussed in Chapter 3.6.

3.3.5. Test Procedure

Before any eye-tracking measurements can be made, the participant must consent to data collection. According to Kröger et al. (2020), eye-tracking data can indirectly contain information about personality traits, emotions, physical and mental health conditions, and more. And since eye activity is not always intentional it may be difficult for users to control what information, they reveal about themselves. This is vital to consider to ensure the integrity of test participants. According to the Swedish Authority for Privacy Protection (2021) biometric data, is considered sensitive personal data and must be protected more than other data. How the personal data is to be processed must be clearly understood by the participant. They must be informed that their data is being collected, why it is collected, and how it is to be used. In addition, they must also be informed about their rights to extract their information or revoke consent.

When using eye-tracking glasses, it is vital to find suitable corrective lens inserts for people with glasses to minimize the risk of them not seeing properly or squinting, thereby causing fragmented data. If a person has a type of sight impairment without corresponding lenses from Tobii it is beneficial to ask them to use contact lenses during the test if possible (Holmes, 2019).

Pernice and Nielsen (2009), suggest that excessive explanation of eye-tracking details to test participants should be avoided as it may lead to unnecessary insecurity. Overthinking what they are looking at could cause participants to exhibit unnatural behavior, which is undesirable. It is nonetheless important to inform participants that the test will involve eye-tracking and to allow them to ask questions about the technology during the test.

Typically, participants perform an initial calibration of the eye tracker by fixating on a pre-defined target, at the beginning of a recording session. Tobii Pro Glasses 3 assists the process by signaling when the calibration is successful (Tobii, n.d. -b).

It is common practice to perform a validation of the calibration by presenting a new set of stimuli for the participant to fixate on. This allows the test leader to evaluate the calibration in conjunction with the test and get a recording of the quality of the eye measurements for later review (Tobii, n.d. -c).

3.4. Eye-Tracking in Design

Despite extensive research on eye-tracking over the past decade, only a small amount of research is focused on how to implement eye-tracking in the design process can be found. Nonetheless, some research conducted in this area has yielded valuable insights that can be further developed. For example, several previous studies state that eye-tracking findings expand on or supplement those obtained through traditional usability (Goldberg et al., 2002; Pretorius et al., 2010; Wang et al., 2019).

Some studies using eye-tracking have been conducted using A/B-testing to find the best version of an interface, in part by comparing the cognitive workload between versions (Zhou et al., 2022). Others have instead evaluated a single interface to determine unnecessary elements or usability issues (Erol Barkana & Açık, 2014; Wang et al., 2019). Fixation duration appears to be one of the most commonly used metrics used in design evaluation-related eye-tracking, showing up in many studies (Zhou et al., 2022; Erol Barkana & Açık, 2014; Goldberg et al., 2002; Wang et al., 2019). These studies also include pupil diameter, number of fixations, dwell times, and saccadic amplitudes to various degrees. AOIs, and metrics related to AOIs, have been shown to work to gain usability insights about specific areas (Erol Barkana & Açık, 2014; Goldberg et al., 2002). Scan paths and heat maps have also proved a valuable tool for examining user behavior, both for comparing different users or user groups, and for seeing if the user behavior corresponds with the intentions of the design (Zhou et al., 2022; Pretorius et al., 2010; Goldberg et al., 2002).

3.5. Test & Usability Theory

This chapter outlines the definition of usability and human factors that should be considered in usability testing, a framework for user test fidelity, and a description of cognitive workload.

3.5.1. Usability

An important goal of human-machine interface evaluation is to effectively characterize the usability of it. Usability is a quality attribute that refers to a system's ease of use. It is defined by the International Organization for Standardization (ISO) 9241 (2018) as 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". Effectiveness, efficiency, and satisfaction are defined as below:

- Effectiveness: "accuracy and completeness with which users achieve specified goals" (International Organization for Standardization [ISO], 2018).
- Efficiency: "resources used in relation to the results achieved" (ISO, 2018).
- Satisfaction: "extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user's needs and expectations" (ISO, 2018).

A good system does not only facilitate ease of use, but it also needs to be functional. Utility refers to whether a system provides the features you need. Usefulness is the combination of usability and utility (Nielsen, 1993).

To evaluate usability, one must understand the user, the product, and their surroundings as a system of interconnected parts (Nielsen, 1993). The definition of usability recognizes that goals need to be achieved and that to do so, tasks need to be performed in a certain context. Thus, usability is not constant. It varies depending on the user, the task, and under what conditions a task must be performed. Therefore, the design of human-computer interaction must consider its usefulness under many different conditions. Conditions that may be hard to mimic in a laboratory. A prerequisite for designing good usability is to understand the user and usage in its real context, an understanding that is acquired from usability tests (Jordan, 1993).

For a user interface to have good usability, the user must be able to understand the current state of the system and their ability to change it (Vermeulen et al., 2013). Gulf of evaluation is a term describing the degree to which a system lets the user perceive and interpret its state. It refers to the distance between a system's response and the user's understanding of that response. Gulf of execution is the degree to which a system allows the user to accomplish a specific goal. It refers to the distance between a user's intention and a system's allowable actions.

Traditional usability methods can provide:

- the time the user spends on a page
- actions the user takes
- what is read aloud by the user
- what the user hovers the mouse over
- smiles or grimaces
- comments (Pernice & Nielsen, 2009)

The three different aspects of usability can be measured in different ways. Below are a few examples for each aspect:

- Effectiveness
    - number of correct tasks
    - number of errors during each task
    - severity level of errors
    - situational awareness achieved using the system
- Efficiency
    - time to complete the task
    - unnecessary actions
    - fatigue

    - time efficiency
    - cost efficiency
- Satisfaction
    - General satisfaction
    - Trust of users
    - Number of users with complaints
    - Evaluation of how the system meets expectations
    - Negative feelings before, during, or after usage (Svensson et al., 2020)

Verbal data from usability tests can be categorized into usability problem types as seen in Table 1.

Table 1

Description of Categories of Usability Problems Found in Verbal Data

Terminology problem(s)	The participant does not understand a term.
Data entry problem(s)	The participant does not know how to enter data.
Comprehensiveness problem(s)	The participant finds that the information is not clear or applicable.
Feedback problem(s)	The feature fails to provide feedback.
Relevance problem(s)	The participant feels that certain information should not be included.
Formulation problem(s)	The participant does not appreciate a formulation.
Visibility problem(s)	The participant fails to spot an element, for example, a link, a button, or information.
Completeness problem(s)	The participant feels that information is missing or more elaboration is needed.
Graphic design	The participant does not appreciate the layout.
Structure problem(s)	The participant finds that the order of information is problematic, or that the structure is not clear.

Note. Categories and descriptions according to van den Haak (2008).

There are many methods used to measure perceived usability. The system usability scale (SUS) was used in this study for post-test assessment of usability and is described below in more detail. A SUS is a ten-item Likert scale for subjective assessment of usability (Sauro, 2018). The result is translated into a SUS score from 0-100. This score is related to an average score of 68. Anything above 68 is considered above average, whereas a score below 68 is considered below average. A corresponding adjective and grade are also given for the different ranges of scores, as a way to describe an experience. The method has a high level of face validity thanks to the variety of usability aspects that the statements cover, including support, training, and complexity (Brooke, 1995). Nonetheless, the SUS is based on self-reported data and measures subjective user perception, not objective performance of the interface, and should be combined with other methods (Laubheimer, 2018). It also does not explain the weaknesses or strengths of the experience, or in what way it can be changed to improve it.

The number of participants that should be used in usability tests is different depending on the objective of the tests. Quantitative tests meant to measure the usability of a system, require larger sample sizes to be useful (Nielsen, 1993). The recommendation is to use more than 20 participants and to check the result by calculating confidence intervals and statistical significance. Qualitative tests, meant to identify usability issues, require fewer participants. Nielsen asserts that five participants are enough to identify 85 percent of the issues in an interface. For usability engineering purposes in general, some data is better than no data.

3.5.2. Jordan's Five Component Framework of Usability

The five-component framework of usability defined by Jordan (1993) describes how a user's performance changes with experience and consists of the following measures:

1. Guessability: "The effectiveness, efficiency and satisfaction with which specified users can complete specified tasks with a particular product *for the first time*" (p.25).
2. Learnability: "The effectiveness, efficiency and satisfaction with which specified users can achieve a competent level of performance on specified tasks with a product, *having already completed those tasks once previously*" (p.26).
3. Experienced user performance (EUP): "The effectiveness, efficiency and satisfaction with which specified *experienced* users can achieve specified tasks with a particular product" (p.27).
4. System potential: "The *optimum level* of effectiveness, efficiency and satisfaction with which it would be possible to complete specified tasks with a product" (p.28).
5. Re-usability: "The effectiveness, efficiency and satisfaction with which specified users can achieve specified tasks with a particular product *after a comparatively long period away from these tasks*" (p.29).

Learnability refers to the novice user's experience in learning to use a system (Nielsen, 1993). Experienced user performance refers to the expert user's experience with an interface once their learning curve has flattened. This breakdown of attributes may lead one to think that an interface is characterized by either learnability or efficiency, however, most complex systems have and need to fulfill both. Despite the categorical distinction between expert and novice users, most expert users are not proficient in all parts of a system. A complex system includes many features and allows for a combination of interactions. They may also be subject to the development of new features, functions, and designs. As a consequence, experts need to seek help for those parts that they are not proficient in, resulting in the need for learnability also in complex user interfaces where the primary users are experts.

The question of expert versus novice users in usability testing has been debated. Nielsen (1993) argues that novice users should typically be used, with expert users included only in specific circumstances. The fundamental principle of usability testing is that participants should represent the actual user base. For complex systems, relying solely on novice users can lead to false negatives, meaning that the testing reveals many usability issues that actual users might not encounter (Redish, 2007). These users would likely encounter different problems. Sauer et al. (2010) found no clear advantage for either user group in their study on user expertise and prototype fidelity. However, they noted that experts might be preferable for identifying issues related to the efficiency of product operation and for addressing problems that go beyond the specific tasks used in usability tests.

3.5.3. Cognitive Workload

Cognitive workload refers to the mental effort required by a user during tasks, influenced by their limited cognitive resources, such as attention and working memory (Chen et al., 2011). Cognitive workload is influenced by many factors, but in particular task load and task design (Longo, 2018). Designers need to consider workload when developing products or designing a task to ensure that the workload is not too high or too low. Low levels of cognitive workload are associated with boredom and a decrease in attention, whereas high workload levels are related to increased error rates and tunnel vision. As a user's cognitive workload increases, their ability to perform effectively decreases until they reach a point of overload, hence cognitive workload and task performance are closely connected in human factors engineering. To measure cognitive workload, a working prototype is necessary to ensure valid measurements. Eye-tracking has the potential to provide a source of objective data on the user's experienced cognitive workload. How eye-tracking can contribute to analyzing cognitive workload will be tested and discussed further in this study.

3.5.4. Test Types

The following is a description of four different types of usability tests: exploratory, comparative, assessment, and verification tests (Homeland Security, 2023). Exploratory tests aim to explore an interface to learn about human-machine interaction and find potential problem areas to study further. Comparative tests are used when comparing two versions of an interface or feature, or to compare between features in, for example, A/B-testing. Assessment tests are used to determine the efficiency or effectiveness of an interface or feature and to locate potential usability problems. Lastly, verification tests are meant to provide objective data to see if the interface passes set requirements (O'Hara et al., 2012).

3.5.5. Four Factor Framework of Contextual Fidelity

The four-factor framework of contextual fidelity considers aspects that influence the realism, or fidelity, of usability testing (Sauer et al., 2010). The level of fidelity on each factor will influence user behavior, hence impacting the results of usability tests. Due to resource constraints, usability test situations may differ from the real usage situation in four ways:

1. The user characteristics of the test participant may be different from those of the real user.
2. The prototype may not be fully functional.
3. The task given is not representative in terms of detail or complexity.
4. The testing environment is different from the natural environment of use.

A more detailed description of each aspect is provided in the following sections.

User Characteristics

When performing user tests, there are a number of user attributes to consider; user attitude, user state, user personality, and user competence (Sauer et al., 2010). User attitude may influence the interaction with a product and can include for example openness to technology and environmental concerns. User state considers temporary conditions that may affect the interaction between a user and a product, for example, fatigue. Different personalities also influence a user's interaction with a product. A thorough person may identify more usability problems for example. Lastly, competence includes a combination of knowledge, skills, and abilities, and is most often of the highest importance in usability testing.

System Prototype

Reduced fidelity prototypes provide the opportunity to gain user insights both faster and cheaper (Sauer et al., 2010). However, the question of how accurate a picture of actual user behavior one gains when using low-fidelity prototypes has been of concern in research. Sauer et al. (2010) discuss the inconsistency that the collection of research on the topic has shown and conclude that current published studies are insufficient in determining a general recommendation regarding the impact of reduced fidelity prototypes

on the accuracy of studied user behavior. However, they state that reduced fidelity prototypes seem to be more effective when usability errors are studied rather than efficiency measures such as task completion time. To determine the level of fidelity of a system prototype the framework distinguishes four aspects to consider: breadth of functions, depth of a function, physical similarity, and similarity of interaction.

Task Scenario

The framework characterizes task scenarios based on two attributes; breadth and depth (Sauer et al., 2010). Breadth refers to the degree to which the complexity of the real task is represented in the task scenario. The depth of a task scenario refers to the level of detail with which a task is performed.

Testing Environment

The testing environment is made up of physical features, social features, and the application domain (Sauer et al., 2010). The physical environment includes stressors such as noise, temperature, and vibration. The social testing environment considers the presence of other humans, for example, the design team. Research has shown that the presence of observers influences user behavior in usability tests. The outcome of usability tests may also be influenced by the domain in which the product is used. For example, the use of a product may differ between a professional setting and a leisure context.

3.6. Think-Aloud

Thinking aloud is a commonly used method in usability research (van den Haak et al., 2003) with Nielsen (1993, p. 195) asserting it "may be the single most valuable usability engineering method." In a thinking-aloud test, the participant is asked to continuously verbalize their thoughts, either while using a system or after having used a system. It enables insight into and reasoning behind user behavior that observation alone cannot provide (van den Haak et al., 2003).

Think-aloud often refers to more specifically concurrent think-aloud (CTA) which is performed during a test. Think-aloud after a test is called retrospective think-aloud (RTA). For the two variants to be equivalent and comparable, an RTA test must include stimuli to enable the participant to recall the task performance (van den Haak et al., 2003). There are advantages and limitations to both methods, however, they both carry the benefit of producing data that reflects the actual use of an artifact and not judgments about its usability. Their strengths include pinpointing problems that may otherwise be forgotten in interviews while providing insight into how users approach tasks and enabling the participant to explain the reason behind an experienced problem. Moreover, the methods are suitable not only for finished prototypes but for lower fidelity prototypes as well (Lewis, 1982) and can be applied with more or less involvement from the evaluator (Alhadreti & Mayhew, 2017).

Although the value of think-aloud is well-established, the validity of RTA and CTA has been questioned. In general, concurrent think-aloud has been criticized for interfering with and not representing automatic cognitive processes (Eger et al., 2007). A test participant is only able to verbally report what they are aware of. Moreover, reactivity, i.e., change in task performance, has been evident in some studies (Russo et al., 1989). Participants may become more methodical and as a result of that perform better, or they may perform worse as a result of a larger workload. One way to minimize the impact of reactivity is to limit the interaction between researcher and participant during a test and allow for uninterrupted think-aloud by the participant (Alhadreti & Mayhew, 2017). Concurrent think-aloud will also disrupt measurements of reaction time, task completion time, scan paths, and heat maps as they are expected to elaborate on their thoughts meanwhile performing the task (Nielsen & Pernice, 2010). The behavior that people may do when they are speaking at the same time as performing a task, which may look like fixations on the eye-tracking data, is called perpetual viewing by Pernice and Nielsen (2009). Moreover, the act of thinking aloud while performing a task may feel unnatural to test participants (Nielsen, 1993) and the verbal reports are likely incomplete as the participant must prioritize between performing a task and thinking aloud (Alhadreti & Mayhew, 2017). While their incompleteness must be considered, the given reports can nonetheless provide sufficient data.

In contrast, RTA allows for an uninterrupted test. Hence, this test will have less reactivity and interference with cognitive processes and allow stopping and reviewing the recording for more details if needed (van den Haak et al., 2003). This has the potential to produce more information from each test, allow for spontaneous and natural think-aloud, and can be particularly valuable when representative users are few (van den Haak et al., 2003; Alshammari et al., 2015). A disadvantage of RTA includes longer test duration, as the test first needs to be performed and then reviewed. A more severe disadvantage regards biased or fragmentary accounts by participants. The participant may forget specific things or conceal thoughts and must also rely on their long-term memory to report their behavior. Cued RTA methods have emerged to minimize these risks. The test participant is presented with a stimulus, e.g., a screenshot or a recording of them performing the task(s) which helps them cue their think-aloud.

Eye-tracking data and cued retrospective think-aloud protocol can be combined in different ways. The video recording produced by the eye-tracking technology can be used on its own as a cue for RTA or with the participant's gaze overlay i.e., gaze-cued retrospective think-aloud (GRTA). When a static stimulus is more suitable, a screenshot can be used or a gaze plot with the participant's eye movements included. However, adding an eye-cue to the traditional retrospective think-aloud method has certain effects that should be considered when adopted in usability testing, such as longer sessions, longer silence periods and the eye-movement overlay may be a distraction to some participants. Elbabour et al. (2017) recommend recruiting participants who are familiar with both think-aloud and eye-tracking to reduce the distraction factor.

Several studies have researched the outcome of the different variants of think-aloud. Hyrskykari et al. (2008) compared the usability findings obtained from concurrent think-aloud to those from gaze-cued retrospective think-aloud, finding that GRTA produces not only more verbal data but also more informative data. The same type of result was shown in a study by Eger et al. (2007) where CTA, RTA, and GRTA were compared. Van den Haak, et al.'s (2003) study indicates that CTA and RTA can be regarded as equivalent in terms of quantitative output. However, out of the usability problems expressed in CTA, more of these could be detected through observation compared to RTA where more expressed usability problems were not observable and could only be detected by means of verbalizations. Moreover, RTA seems to be less influenced by task difficulty. Difficult tasks risk causing reactivity and incomplete verbalizations when performing concurrent think-aloud. Their study also showed that RTA led participants to give explanations and suggestions behind their actions to a greater degree compared to participants that performed CTA which tended to give more descriptions of their actions. A study by Elbabour et al. (2017) comparing two different cues in RTA found that the participants performing gaze-cued RTA produced more comments regarding their perception e.g. "I had a little look around to see what is for sale" compared to video-cued RTA where the participant produced more comments regarding their manipulation e.g. "I clicked on this link". Olsen et al. (2010) compared no-cue RTA, video-cued RTA, gaze-plot-cued RTA, and gaze-cued RTA and concluded that any kind of cue was shown to produce more verbalizations, but with little difference between what type of cue was used. It should be stated that none of these studies were performed on complex user interfaces.

Below is a non-exhaustive list of tips for performing think-aloud protocols for practitioners presented by Pernice and Nielsen (2009):

- Do not perform gaze-cued retrospective think-aloud when you have eye-tracking tests left to complete. It may cause the participant to become too aware of their gaze.
- In GRTA, make sure to inform participants that rapid eye movement is normal.
- Perform RTA right after the performed task to maximize the participant's ability to recall information.
- Instructions should be given in a way that does not invite introspective explanations. The participant should be asked to report their thoughts, not explain them.
- Display a snippet of the gaze-cued stimuli before GRTA to allow the participant to become familiar with seeing their gaze.

3.7. Interpretation of Eye-Tracking Data

Eye-tracking devices output a large amount of data, with Tobii Pro Lab allowing for 89 different metric visualizations, in addition to the raw data export (Tobii, 2024b). To narrow down which data is important, and how it can be interpreted, previous research can be leveraged.

3.7.1. Areas of Interest

A common tool to deepen the insights from eye-tracking metrics is areas of interest (AOIs). An AOI is a region in the stimulus that is defined and studied more closely (Holmqvist et al., 2011). By defining an AOI it is possible to gather data about specific elements in the stimulus, instead of looking at the data as a whole. AOIs can be both static and dynamic depending on whether the stimuli are moving or not. Regardless of type, it is important to define AOIs before conducting tests, since changing the AOIs after gathering the result would entail changing the hypothesis. The hypothesis should guide the placement of the AOIs, as these are inevitably interconnected. Holmqvist et al. (2011) propose the concept of expert-defined AOIs, which means having an expert assist in determining how to define relevant AOIs. Air traffic control is presented as an example of this use, in which an air traffic controller could guide relevant placements.

Holmqvist et al. (2011) also present several guidelines for using AOIs:

- Avoid overlapping AOIs to avoid overlap in the data.
- Each AOI should only contain semantically similar elements.
- AOI size is limited by the accuracy of the data. A higher accuracy allows for smaller AOIs.
- Do not use too many AOIs, only as many as the hypothesis demands.

Through the use of AOIs, it is possible to get AOI-related metrics (Tobii, 2024b). These include the same metrics that would otherwise be used for the whole stimuli, presented later in this chapter. However, some AOI-specific metrics also exist, such as:

- Time to first fixation. Shows the time it took until the participant placed its first fixation within the AOI.
- Duration of glance. Shows how long the participant spent looking at the AOI before leaving. This metric can be presented as individual glances, an average, or a total of all glances.
- Number of glances. Shows how many separate times the participant looked at the AOI, having looked at something else in-between.

These metrics can be used to compare the attention-grabbing and attention-holding properties of AOIs. It is also possible to use these metrics to analyze the data in other ways depending on the hypothesis and context, such as determining which AOIs were glanced at, in which order the AOIs were glanced at, or which AOI was glanced at the most (Holmqvist et al., 2011).

3.7.2. Metrics

Eye-tracking metrics are quantitative data that can be the subject of statistical analysis (Holmqvist et al., 2011). Which data, and how it is analyzed are determined by the hypothesis and goal of the research. One such goal could be to compare the cognitive workload between different tasks or versions of an interface (Yang et al., 2014; Goldberg & Kotval, 1999). If eye-tracking metrics can be used to determine cognitive workload has been the subject of many studies (Ehmke & Wilson, 2007; Kovesdi et al., 2015). While these studies have shown that cognitive workload can be determined by eye-tracking metrics, most of these have been conducted in lab environments with limited and specifically prepared stimuli. Further, because variations in task, stimuli, environment, and subject can all influence the metric data it is usually not helpful to quantify which value ranges are high and low (Zagermann et al., 2016). Each test participant will have a different baseline for some metrics and thus cannot be compared directly (Iqbal et al., 2005). It is also important to note that the metrics will never reveal why the data indicates something, only that it does (Stieff et al., 2011).

Fixation Metrics

The use of fixation-related metrics is a common way of analyzing eye-tracking data (Borys & Plechawska-Wójcik, 2017). Since fixations are the instances where visual information is acquired, it can be used to indicate differences in cognitive workload or potential usability issues (Holmqvist et al., 2011; Goldberg & Kotval, 1999; Jacob & Karn, 2003). Fixation metric indicators are:

- Higher number of fixations. Can indicate less efficient search (Goldberg & Kotval, 1999) or a higher cognitive workload, specifically a difficulty processing information quickly (Wang et al., 2014). However, depending on the context and definition of cognitive workload, fixation duration has also been shown to decrease with increased load, for example flying and driving (Skaramagkas et al., 2023). This is possibly due to a high perceptual load causing insufficient capacity to process irrelevant stimuli (Wang et al., 2014).
- Low fixation rate. Can indicate a higher cognitive workload (Chen et al., 2011).
- Longer duration of fixations. Can indicate a higher cognitive workload (Skaramagkas et al., 2023).
- Higher number of fixations in an AOI. Can indicate that the area is more noticeable or contains important elements to the user (Jacob & Karn, 2003).
- Longer total duration of fixations in an AOI. Can indicate that the user requires a lot of processing in a specified area (Cowen et al., 2002).
- Longer time to first fixation in an AOI. Can indicate that that an element has low noticeability (Bergstrom & Schall, 2014; Holmqvist et al., 2011).

Saccade Metrics

Similar to fixation metrics, saccade metrics are commonly used in eye-tracking. In contrast to fixation metrics, saccades hold information about the searching and movement part of the gaze data (Borys & Plechawska-Wójcik, 2017). Even if visual information is not actively processed during saccades, it can reveal some information about the cognitive state of the user (Nielsen & Pernice, 2010; Zagermann et al., 2016). Saccade metric indicators are:

- Lower saccade amplitude. Can indicate that trajectories are less pre-planned and insufficient visual clues in the interface (Goldberg & Kotval, 1999; Goldberg et al., 2002). Can also indicate a higher cognitive workload (Keskin et al., 2020; Van Orden et al., 2001).
- Lower peak saccade velocity. Can indicate a higher cognitive workload. Correlates with saccade amplitude (Di Stasi et al., 2011).

Pupil Dilation Metrics

While much research has stated that pupil dilation metrics can be used to indicate the cognitive workload of a user (Chen et al., 2011; Porta et al., 2012; Klingner et al., 2008), it is a difficult metric to use. Pupils are not only affected by the current task but by the amount of light coming from the surroundings, which may not always be constant. Pupil size changes due to environmental factors are much larger than those incurred by task difficulty (Carter & Luke, 2020). Pupil dilation also varies greatly between people in the same conditions. As such these metrics can only be used in highly controlled environments and cannot be compared between different subjects. Nonetheless, under the right conditions, it can be used to infer cognitive workload. The main pupil metric for indicating cognitive workload is:

- Larger average pupil diameter. Can indicate an increased cognitive workload (Chen et al., 2011; Porta et al., 2012; Klingner et al., 2008).

Ratios and Calculations

Some inferential measures can be calculated using the base metrics given as data output to gain further insights. Fixation to Importance Ratio (FIR) is one such measure that looks at how much relative time the user spends with their gaze in the most vital areas (Kovesdi et al., 2015). Selective Attention Effectiveness (SAE) expands further on this result using the FIR to calculate a single number representing the user's ability to focus on what is most important.

Fixation/Saccade Ratio is a potential way of determining if more information processing or search activity was required for a task proposed by Goldberg and Kotval (1999). This ratio is given by dividing the total duration of fixations by the total duration of saccades in a time of interest. While the study by Goldberg and Kotval (1999) did not yield any significant difference between interfaces, it is possible to use this ratio with the purpose of finding differences in the amount of processing and search.

Percent Change of Pupil Size (PCPS) is a way of countering the otherwise very subjective dilation data. Instead of looking at just the diameter, this measure instead looks at the change from a baseline measurement (Iqbal et al., 2005). A larger percent change in pupil size (PCPS) can indicate increased cognitive workload. This is a measure that compares the change to a baseline measurement. This is calculated as (Measured Pupil Size - Baseline Pupil Size)/Baseline Pupil Size.

Average Percent Change of Pupil Size (APCPS) is a single value calculated by taking the average PCPS over a specific time period.

3.7.3. Visualizations

In addition to the metrics used to analyze eye-tracking data, visual presentations of the data are commonly created. These visualizations offer a way to overview the result without statistical analysis but are also a way to analyze other aspects than those found in the metric data. More than with metrics, however, these are open to subjective interpretations.

Scan Paths

A scan path is one of the ways in which eye-tracking data can be visualized. These can be defined as a sequence of alternating fixations and saccades visualized on a display (Goldberg & Kotval, 1999). Holmqvist, et al. (2011) state that previous research conducted on the area of eye-tracking and scan paths has found it difficult to find ways of relating scan paths to the cognitive process of the user and that scan paths have generally proved to be difficult to analyze in a way which would yield objective data. The authors do present ways of statistically comparing scan path likeness but also state that this does not provide much insight into real-world usability problems. Instead, scan paths can be used in manual analysis to find potential issues to investigate further, to control the quality of the data, as a simple visualization to convey otherwise complex data, or as a mediating tool when conducting a retrospective analysis together with a user (Holmqvist et al., 2011). In one study, looking at how eye-tracking metrics relate to usability, the authors propose how some scan path characteristics can indicate problems with usability (Goldberg & Kotval, 1999). While these do not provide an objective result, nor indicate the cognitive process of the user, they can be used as guides to know what to look for when analyzing scan paths for potential usability issues. Indicators for potential issues are:

- Irregular scan paths. Can indicate that the user deviates from a regular scan cycle (Goldberg & Kotval, 1999).
- Large spatial density (spread of fixation points). Can indicate that the user has an indirect search approach and does not know where to look (Goldberg & Kotval, 1999).
- High frequency of transition between AOIs. Can indicate that the user has an inefficient search (Goldberg et al., 2002).

Scan transition matrices are one way to systematically analyze scan paths (Ponsoda et al., 1995). By expressing the number of transitions between individual AOIs, it is possible to gain information about the search pattern (Goldberg & Kotval, 1999). The density of this matrix can be calculated as a single value for comparison.

Heat Maps

A heat map is a way to visually represent the spatial distribution of data (Holmqvist et al., 2011). While heat map visualizations may not be as objective as statistics, they can serve as a simple and useful tool for communicating conclusions to clients or stakeholders.

Even though a heatmap can be useful in identifying areas that people looked at more, it is difficult to draw any conclusions about why they looked there (Holmqvist et al., 2011). It could be because of some attention-drawing element, but can also be due to confusion, and needing more time to understand the information. It is therefore reasonable to use heatmaps to determine these points of interest but then use other methods to determine the cause of the behavior.

If conclusions are to be drawn from heat maps, Nielsen and Pernice (2010), recommend having at least 30 participants so that individual behaviors can be generalized and the influence of outliers reduced.

When creating heat maps, a wide range of settings are usually available to specify the visual look of the picture. To allow for comparisons, it is important to keep these settings consistent between pictures (Holmqvist et al., 2011).

04

General Process & Execution

This chapter encompasses the overarching process of the study, including a general walk-through of the two testing phases and a description of how the resulting C2ET method was developed. User tests and their fidelity, experiment design, and materials are also presented.

4.1. Overarching Process

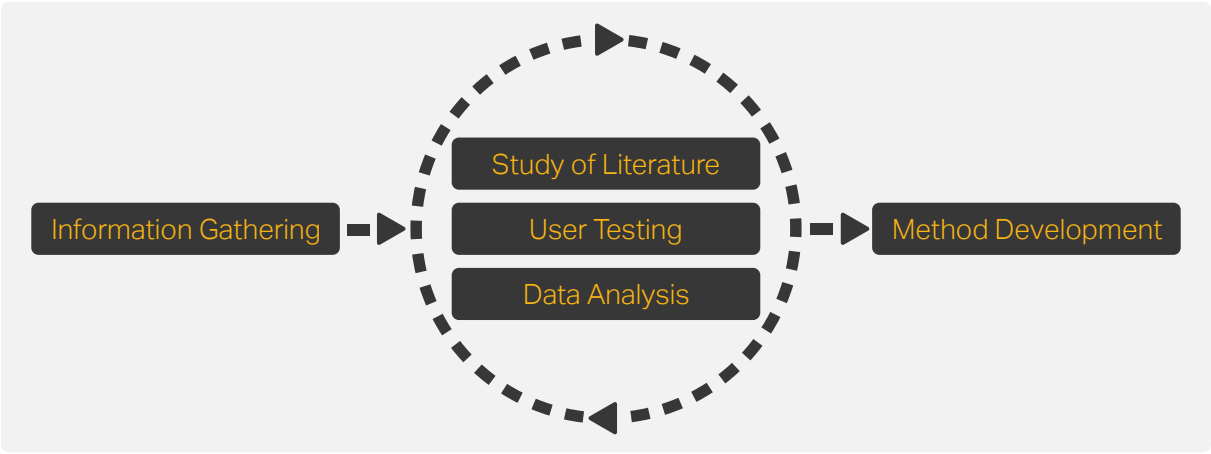
Due to the lack of previous research on the use of eye-tracking for usability evaluation of C2 interfaces, this study employed an exploratory approach. By reviewing existing research on usability, user testing, and eye-tracking, and testing the theories in the context of Saab, an incrementally increasing understanding of the limitations and possibilities of using eye-tracking to evaluate C2 interfaces, was acquired. The aim was not to provide conclusive solutions or final answers to how eye-tracking is to be implemented but to add to the body of eye-tracking research with a contextual method for the evaluation of C2 interfaces in a corporate environment. The exploratory approach gave way to a flexible way of working where learnings were continuously incorporated furthering the work. The HFI team at Saab Surveillance was regularly consulted throughout to ensure that the project resulted in a useful and relevant method for their work.

The general process of the project is illustrated linearly in Figure 5 with an iterative process in the middle. The work started off with gathering information about various topics; eye-tracking technology, eye physiology, eye-mind hypothesis, the history of eye-tracking, and the use of eye-tracking in usability research, presented in Chapter 3. The eye-tracking glasses were also trialed along with the associated software Tobii Pro Lab. After familiarizing with the technology and current research the process of designing user tests began. This was done first by consulting research on the topic of usability testing, eye-tracking tests, think-aloud protocol, and eye-tracking data interpretation. The next step was to prepare and perform user tests. Two iterations of user tests were performed, where learnings from the first were transferred and developed in the second. Test A and B in testing phase one, and C and D in testing phase two. The data obtained from both phases was analyzed and interpreted using various evaluation methods

resulting in a collection of qualitative and quantitative findings on the usability of the C2 interface. The learnings from the procedure of performing eye-tracking tests and data analysis were collected and formed the resulting C2ET method, a method for the use of eye-tracking in the evaluation of C2 interfaces.

Figure 5

Visualization of the General Process of the Study



Note. Beginning with information gathering, iterating through the testing phases, and then concluding with the method development.

The steps of each iteration of user tests are described in a process image in Figure 6. Previous research on usability testing and eye-tracking applications, in combination with knowledge of Saab's way of working, and the direction from research questions created the basis for formulating hypotheses. The hypotheses related test design factors, including user expertise, prototype fidelity, and level of interaction, to the application of eye-tracking technology and the resulting eye-tracking data. Based on the hypotheses, experiments were designed, and tasks formed. At this stage, AOIs were also defined, and which metrics were to be analyzed was decided. Operators at Saab guided the design of appropriate test tasks. The test setup was then piloted, and certain adjustments were made before user tests were performed. This process was repeated in the second iteration of tests while incorporating the learnings from phase one.

Figure 6

Visualization of the Linear Process Within Each Phase Iteration

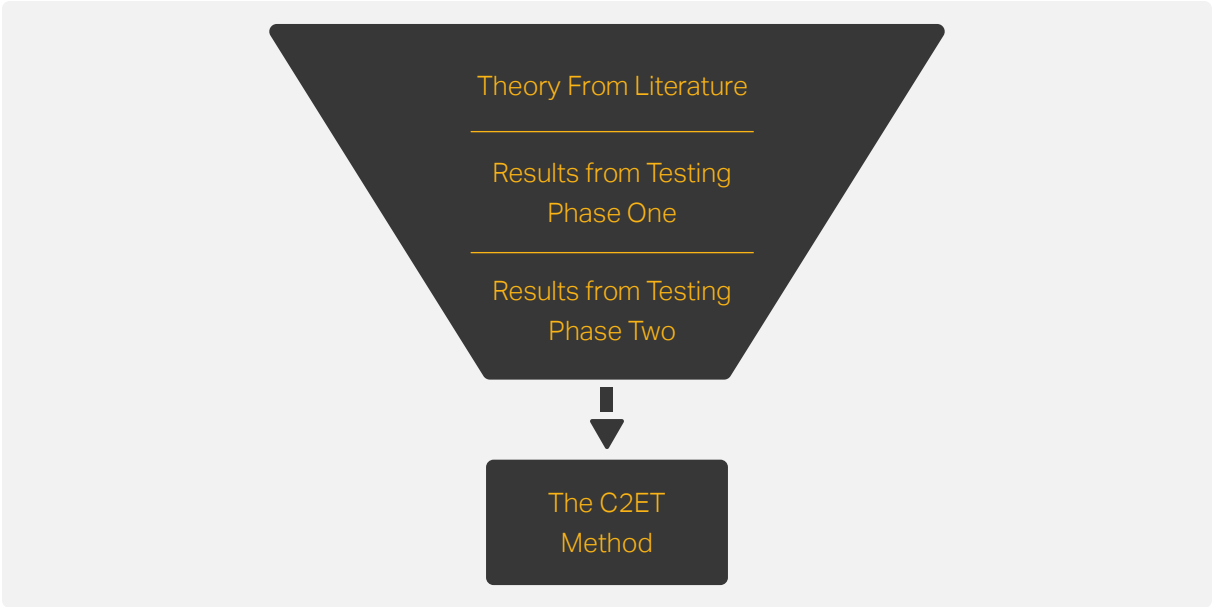
The first testing phase aimed to assess how two design parameters, prototype fidelity and user expertise, affect the usefulness of the eye-tracking data. This was done by performing two tests on test objects with contrasting levels of system prototype fidelity, where reaction time was measured and the overall interaction with the interface was observed. Test participants with varied levels of experience with the C2 interface were included to study how expertise impacts the usability findings derived from eye-tracking data.

The second testing phase aimed to see how the level of interaction with the interface influenced the possibility of deriving usability insights and measuring cognitive workload from the eye-tracking data. Auxiliary methods to eye-tracking, including gaze-cued retrospective think-aloud and a usability scale, were also trialed to see how eye-tracking can be triangulated with other methods. Two tests were performed, the first with a limited test scope and restricted level of interaction with the interface. The second with a wider test scope and unrestricted level of interaction.

The process of performing user tests and the result of eye-tracking data analysis resulted in the C2ET method. A process that can be described as a funnel characterized by a wide opening that narrows down towards the bottom (see Figure 7), illustrating going from broad open-ended questions to narrowly scoped questions and ultimately, answers. The research questions stated in Chapter 1, provided the overall direction of the study while hypotheses were formulated as a framework for answering said questions. The hypotheses were evaluated by combining current research on eye-tracking and usability with the results of user tests. User tests were necessary to determine whether the research was applicable to the context of Saab and to identify what contextual adjustments were needed when applying the technology to dynamic interfaces such as C2 interfaces. For a method to be relevant and applicable, it needs to guide its user. In other words, the method is intended to be general enough to apply to different contexts within Saab, while specialized enough to provide necessary information. More on how the method was developed will be presented in the next section.

Figure 7

Visualization of the Process by Which the C2ET Method was Created



Note. Represented as a narrowing funnel as more knowledge is gained by different means.

4.2. Method Development

As previously mentioned, findings from research along with practical experience from user tests formed the C2ET method developed in this study. The basis for how to perform user tests with eye-tracking was found in previous eye-tracking studies and theories on user testing. Thus, the method is influenced by these sources in combination with the findings from testing phases one and two that were compiled together with the practical learnings of both.

The aim was to create a method that was both explanatory and practical. During this study, it became apparent that guidelines for how to plan and execute eye-tracking studies as well as how to analyze and interpret the results would be the essential parts of the method. What the practitioner's objectives are and what type of data they want to receive, influences the design of eye-tracking tests. Therefore, it was decided that the structure of the C2ET method should align with the order of decisions that must be made. For example, the type of test one decides to perform influences the number of participants one should include, and the appropriate evaluation methods. The content of the C2ET method was therefore structured in three sections: test design, evaluation methods, and preparations and procedure. The aim was that the eye-tracking practitioner be guided through the different aspects that need to be considered in the right order.

The content of the method concerning test design was mostly based on the learnings of performing user tests. Recommendations regarding the environment and participant criteria are guidelines provided by Tobii, the producer of the eye-tracking glasses used in this study. Four different test types are presented along with their purpose, a recommendation for main evaluation methods and possible secondary evaluation methods as well as limitations of the test type. Test length is based on constraints of the eye-tracking technology, certain evaluation methods, and general usability recommendations from literature. The recommendations on a number of test participants come from a combination of usability theory and practical experience where this study has shown that usability findings can at times be acquired with fewer participants than theory prescribes. Prototype fidelity is mainly based on practical experience of using eye-tracking on dynamic interfaces which has proven to be less straightforward compared to static interfaces.

The guidelines for evaluation methods include recommendations on how they should be conducted, including steps to perform. In the previous section, appropriate evaluation methods are listed for the different test types. In this section, the practitioner can learn more about them. The relevance of the evaluation methods that have been tested was found in previous research. And the recommendations for how to effectively apply them were mostly gained from the user tests in this study.

Preparations and procedure include a number of practical preparations that should be done before conducting an eye-tracking test, a materials list, and an ordered list of steps that should be taken before and during an eye-tracking test. The list of steps serves as a summary of the topics discussed more thoroughly earlier in the method and can be used as a checklist. The content in this section is based on the collection of practical learnings that have been obtained during the study along with tips from the eye-tracking supplier and previous research.

General recommendations on how to perform usability testing were excluded from the C2ET method. The practitioner is assumed to have general knowledge within the area of usability and user testing.

4.3. Fidelity

The fidelity of the user tests in this study is described in this section based on the Four Factor Framework of Contextual Fidelity as outlined in Chapter 3.3.6.

4.3.1. User Characteristics

User competence was of high significance in the selection of test participants. The terms expert and non-expert were used to categorize test participants. Expert signifies a participant who has been trained in the interface. Non-expert means the participant has had limited training in the interface and no real-world experience in operating it. The term non-expert is used instead of novice, as these participants have some experience with the interface from a designer point of view. It was deemed relevant to perform tests on experts as these are the true users of the interface and their expertise must be facilitated by the interface on an expert user performance level as well as with regards to learnability. However, with limited availability to experts in the specified context, non-experts were used to see if and how they could be included in user tests with eye-tracking.

The other user characteristics: attitude, state, and personality, listed in the Four Framework of Contextual Fidelity (see Chapter 3.3.6), were not deemed relevant in this study and thus not considered in the selection of test participants.

4.3.2. System Prototype

System prototype fidelity, and all four aspects of it, was considered in the tests of this study. It was of particular interest in testing phase one where it was the object of research, see Chapter 3.6.6 for more details. The majority of tests were performed on a system prototype of high fidelity in which all four factors: breadth of functions, depth of function, physical similarity, and similarity of interaction, were of a high degree. One test was performed on a low-fidelity prototype. In this case, mostly visual similarity of the interface was taken into account.

4.3.3. Task Scenario

Overall, the complexity of real-use tasks was not easily mirrored in a lab environment. One important example to mention is the lack of communication in the tests, both within the crew and with other actors that otherwise can constitute a large part of the work of an operator and their cognitive workload. Aside from communication, the tasks in most tests were of high fidelity, although at varying levels. Two tests included an initial mission brief to mimic a real-use scenario.

4.3.4. Testing Environment

The testing environment in which the human-machine interaction took place during the tests is considered to be of low fidelity compared to the real-use environment. The user tests were performed in a lab environment with considerably lower levels of physical stressors such as noise and vibrations. Due to technical constraints, the eye-tracking glasses require the presence of a test leader in close vicinity to the participant during the test. Furthermore, the glasses, although less intrusive than other eye-tracking technology, add another aspect of unnaturalness to the testing environment.

4.4. Materials

All tests were conducted using Tobii Pro Glasses 3 connected with an ethernet cable to a computer. The eye-tracking glasses record sound, video, gaze, and other eye data. The interfaces with which the participants interacted were a C2 interface and a mockup of a C2 interface on a laptop connected to a larger monitor. An audio recorder was used in all steps of the tests to collect voice recordings and the sound of mouse clicks. The GRTA was recorded with a video camera to be able to hear the participant's verbalizations and see the accompanying stimuli. The stimuli for GRTA were projected on either a computer monitor or a projector screen. Data was analyzed using Tobii Pro Lab (version 1.232.52758) and MATLAB (R2023b).

05

Testing Phase One

This chapter accounts for the execution and the results of tests A and B in testing phase one. Discussion about the findings and learnings from this phase is also included.

5.1. Execution

This section contains a description of the test objectives and hypotheses of testing phase one as well as specifics about the test design including participants, test objects, and tasks.

5.1.1. Test Objectives & Hypotheses

Testing phase one aimed to assess how two design parameters, prototype fidelity, and user expertise, affect the usefulness of the eye-tracking data and how usability tests using eye-tracking can be designed efficiently while producing valid results. The hypotheses of testing phase one were:

1. Eye-tracking data from a low-fidelity system prototype is comparable to data from a high-fidelity system prototype of a C2 interface.
2. Non-expert eye-tracking data can be used as a substitute for expert data.

If it is possible to implement the use of eye-tracking on prototypes of lower fidelities and gain results that are transferrable to the real product, then that would allow for testing earlier in the development process when changes are still possible. If it is also possible to use non-experts during testing, the prevalence of testing could increase due to less reliance on the limited number of proficient operators employed by the company.

With regard to prototype fidelity, the aim was twofold. Firstly, to determine whether eye-tracking data on a low-fidelity prototype is comparable to that of a high-fidelity prototype. Secondly, to see whether A/B-testing can be performed on low-fidelity prototypes and whether a difference in usability can be distinguished with the use of eye-tracking.

5.1.2. Participants

Five participants were recruited for phase one, all of which were employed by Saab and familiar with the C2 interface. Three were non-experts, all women, and two were experts, both men. Both expert users had a background in system-related military service and rated their subjective level of expertise as four and five respectively. No participants had undergone eye surgery or had a current medical eye condition other than impaired vision that could be compensated for with prescription glasses.

5.1.3. Test Object, Test Design & Tasks

Testing phase one included tests A and B, performed by all participants and conducted back-to-back. The test object for test A was a low-fidelity mockup of the C2 interface. The mockup was visually similar to the C2 interface but limited in all other regards: breadth of functions, depth of function, and similarity of interaction. It consisted of generic targets scattered on a map. The test object for test B was a real simulator for a C2 system. The C2 simulator is of high system prototype fidelity and is similar to the real user interface. To be able to test something specific, it was decided that reaction times to notifications would be measured. Since identifying targets is a common task for operators, this was included as a secondary task to make for a wider task scenario. Due to the collection of eye-tracking data, where every participant has an individual baseline, the tests in testing phase one had a within-subject design where all participants were exposed to all test tasks.

Before each testing session, the participants were informed about the purpose of the test, what type of data would be collected, and what it would be used for. The participants also filled out a form consenting to the collection and storage of their data in accordance with the EU's General Data Protection Regulation (GDPR). The same form also included a survey asking about general information such as gender, year of birth, familiarity with the interface, and system-related military experience. Moreover, they were asked whether they had a current medical eye condition aside from impaired vision and if they had undergone eye surgery. The full questionnaire can be found in Appendix A.

Test A

The tasks in test A were of low task fidelity with limited and specified interaction. The participant was first instructed to find and click on notifications, unknown targets, and colored buttons. They were to prioritize notifications before targets and buttons.

After this, calibration of the eye-tracking glasses was performed. In line with the recommended procedure from Tobii (Tobii, n.d. -b), the participant was asked to hold up a high-contrast card until the test leader received confirmation from the Tobii Pro Glasses 3 software that the calibration was successful. The process was repeated 2-3 times to ensure valid calibration. The next step was to start the recording and to validate the calibration. The participant was asked to fix their eyes on a target on a slide show as it moved on the page. This step provided the test leader with confirmation that the calibration just performed had been effective and that the eye-tracking data about to be recorded would be accurate. It also produced a recording of the quality of the calibration.

During the test, an unknown target appeared 90 times, one button was colored on 15 occasions, and two different designs of notifications appeared on the screen in randomized order a total of 12 times, 6 for each design. One of the notification designs was a replica of the current notifications in the real C2 interface, with which participants were familiar. The other design was a new concept, which was not familiar to the participants.

Test B

In test B the participants first received a mission brief from the test leader explaining the scenario, decision, and command, a procedure commonly used in military contexts. The participants were tasked with using the C2 interface to detect targets based on certain criteria and identifying these as hostile. They had no specified limitation in their interaction with the interface.

After this, as in test A, calibration and validation were performed in conjunction with the start of the recording. During the test, a number of notifications were activated sporadically. Due to test technical issues in the system simulator, the same number of notifications could not be achieved there as in the low-fidelity simulation. After 20 minutes the test was ended. This task scenario mimicked a real-use scenario to a greater extent than test A, as it involved a mission brief, greater freedom in interaction, and more problem-solving.

Post-Test

After one test, a pilot trial of a gaze-cued retrospective think-aloud (GRTA) was conducted. The eye-tracking recording was played back to the participant as a stimulus, and they were asked to verbalize their thoughts based on their memory of the test and what they saw from the recording. The GRTA was recorded using an audio recorder and video camera directed towards the screen to be able to connect verbalizations to the eye-tracking video recording that was viewed by the participant.

5.1.4. Analysis of Results

The most important metric for measuring the reaction time to the notifications was *time to first fixation*. This time was to be measured from when the notification window first appeared, so two preparations had to be made in Tobii Pro Lab. Firstly, times of interest (TOIs) had to be created. These started at the moment that the notification window was visible and ended as the participant had interacted with the window or after a substantial amount of time had passed. Secondly, areas of interest (AOIs) had to be created which covered the notification windows for the entire TOIs. Since the AOIs had to be placed on the recording, and not on a mapped still frame, dynamic AOIs had to be used. Creating the dynamic AOIs entailed creating keyframes for each time the participant's head moved. This was done for all notifications in all recordings.

With the TOIs and AOIs completed, each time to first fixation could be extracted and labeled with participant and test. Any completely missed notifications were also noted. This data was then divided into three different comparisons: System simulator versus low-fidelity mockup of current notification design, new versus current notification design in the low-fidelity mockup, and experts versus non-experts for both prototype fidelities. Means, medians, and standard deviations were calculated and compiled for each category.

To determine the cognitive workload of the participants the following metric data was analyzed, based on the theory presented in Chapter 3.7:

- fixation duration
- saccade amplitude
- peak saccade velocity
- pupil diameter

Fixation rate was not included in the analysis since the inclusion of fixation duration provides the same insight. A decision was made to only consider long fixation durations as high cognitive workload to simplify the analysis, as these seemed most applicable to C2 interfaces.

Separately from the use of eye-tracking metrics, the recordings were studied with gaze overlay, i.e., gaze replays, for qualitative analysis. This was done by playing the recording inside of Tobii Pro Lab and noting down times when potential usability problems could be identified. The GRTA was transcribed, and relevant quotes were extracted that identified either usability issues with the interface, or how well the GRTA worked in practice. These findings were then compiled into lessons about GRTA as a tool. This, together with the subjective experience of the authors performing the analysis, was then combined into the general findings of phase one.

5.2. Results

This section presents the results of testing phase one. For these tests, the gaze sampling quality was between 90 and 99 percent for all recordings. The lack of statistical significance in the presented data can mainly be attributed to the low number of test participants, but also the exploratory nature of the study. Data in this chapter is therefore only used as an indication and suggestion about which direction to proceed.

5.2.1. System Prototype Fidelity Comparison

Due to some high outliers in the data and the small sample size, the mean value becomes somewhat skewed. Instead, looking at the median value gives a more accurate picture of the participants' performance (see Table 2). The data shows that the participants' reaction time was slower in the system simulator, indicating that the given subtask of identifying tracks and clicking on them, in a low-fidelity prototype did not do enough to mimic the simulated system.

Table 2

Time to First Fixation Data from Tests A and B

	System simulator, current notification		Low-fidelity, current notification	
	Expert	**Non-expert**	**Expert**	**Non-expert**
Nr of data points	14	4	9	18
Mean (ms)	543	1444	323	456
Median (ms)	501	1236	230	230
Standard deviation (ms)	169	1027	187	438

Note. The data is separated by interface type (system simulator and low-fidelity mockup) and participant category (expert and non-expert). Instances when the participant missed the notification completely are not included in the data.

Comparing the cognitive workload of the two fidelities, the data suggests that the system simulator required a somewhat higher cognitive workload than the mockup (see Table 3). However, the difference seems more substantial for experts. The pupil diameter data disagree with the other metrics, pointing towards the opposite. However, because the two test types took place in different environments, there is a high likelihood that environmental factors affected the data more for this metric. It is also clear when studying the recording, that instances of high pupil diameter often occur when the participant looks off-screen. Regardless, the suggestion is that the low-fidelity mockup was not able to mimic the cognitive workload achieved by the system simulator.

Table 3

Cognitive Workload Metrics for Tests A and B

	System simulator		Low-fidelity mockup	
	Expert	**Non-expert**	**Expert**	**Non-expert**
Fixation duration (ms)	585 (894)	541 (565)	357 (391)	483 (582)
Pupil diameter (mm)	3.07 (0.28)	2.64 (0.18)	3.47 (0.19)	2.92 (0.30)
Peak saccade velocity (°/s)	264 (222)	253 (169)	308 (159)	264 (176)
Saccade amplitude (°)	6.42 (4.9)	6.9 (4.5)	9.14 (5.4)	7.6 (5.3)

Note. Data from tests A and B showing eye-tracking metrics tied to cognitive workload separated by interface type (system simulator and low-fidelity mockup) and participant category (expert and non-expert). Each metric is accompanied by its mean value and standard deviation, presented in parentheses. The mean value reflects the average means among participants within their respective categories, while the standard deviation indicates the average standard deviation among participants within each category. The values represent averages for the entire test. Fixation duration and pupil diameter positively correlate with cognitive workload, while peak saccade velocity and saccade amplitude negatively correlate with cognitive workload.

Comparing the two notification types, the current notification had a lower mean and median value compared to the new concept (see Table 4). While the reasons for this could be many, such as participant familiarity with the current version, it does indicate the possibility of differentiating between concepts in a low prototype fidelity stage of development. Due to the small number of participants, it is difficult to say how much familiarity played a role in giving the current version a quicker reaction time. Looking only at the two expert operators, who should be the most familiar with the current version, the result is inconclusive. One operator performed slightly better with the new concept (Current: 543 ms, Concept: 412 ms), while the other performed almost five times worse (Current: 210 ms, Concept: 1031 ms).

Table 4

Time to First Fixation Data from Test A

	Low-fidelity, current notification		Low-fidelity, concept notification	
	Expert	Non-expert	Expert	Non-expert
Nr of data points	9	18	11	17
Mean (ms)	323	456	642	488
Median (ms)	230	230	230	371
Standard deviation (ms)	187	438	378	349

Note. The data is separated by notification type (current and new concept) and participant category (expert and non-expert). Instances when the participant missed the notification completely are not included in the data.

5.2.2. Experience Level Comparison

When testing reaction times to notifications, despite few data points for the non-experts, the data shows a quicker reaction time for experts (see Table 2). A low sample size is present in all data but is especially apparent in this one. While reaction time to notification is a very specific measure and may not be generalizable for other metrics, there is at least some indication that participant of different experience levels performs differently.

When studying the recording of each respective participant, there is also a clear difference between how expert operators and non-experts approached and executed a task. Familiarity with the interface leads experts to utilize more of the interface functionality, something that would be especially important in exploratory tests.

Table 3 shows that the metrics indicating cognitive workload were similar when testing on the system simulator for both experts and non-experts. For the low-fidelity mockup, the results were also similar, but with a slightly larger difference. Importantly, the differences all point in the same direction for experts and non-experts, indicating some consistency. Another interesting observation is the difference in consistency when it comes to fixation duration standard deviation, with experts' fixation durations varying more in the system simulator test. Whether this is due to individual factors, or a characteristic of experts is difficult to say. Overall, the results seem inconclusive, with the results not being able to determine if cognitive workload data differs between experts and non-experts.

5.2.3. General findings

In addition to the insights gained by the comparisons of metric data previously presented, other findings emerged through qualitative analysis. During viewings of gaze replays, general usability problems of the interface could be identified. One such example is when a participant attempted to perform an action that was not available and failed to understand why it did not work, possibly due to a lack of feedback. Other findings which were possible to detect through the use of eye-tracking, were the instances where the participants looked at the notification, ignored it, and continued working in the interface. Using a screen recorder, it would have been impossible to tell if the participants had acknowledged the notification in this case.

During the analysis, a lot of time was spent manually tracking the dynamic AOIs onto the notification windows. Even though only about 3 seconds were tracked for each notification, this resulted in a substantial amount of work. Dynamic AOIs, while an important feature, are therefore very time-consuming. The alternative, assisted mapping of the entire screen, also comes with issues. If done for the entire session, the AOIs placed on the mapping do not follow for example if a window is moved. Assisted mapping was also seldom completely reliable when used in a setting where the interfaced changed appearance. This makes it harder to use for AOI purposes, especially smaller ones but still performs well for producing heat maps. If the recording is instead segmented into different images every time something is moved and mapped separately, this is instead very time-consuming.

From the single GRTA that was conducted in phase one, several lessons were learned. The first was the need for a clear objective with the think-aloud. The participant was left to verbalize whatever thoughts they had as no other goal had been set with the think-aloud. It resulted in the participant evaluating their performance rather than providing reasoning behind decisions and actions. The lack of objective also made for a less fruitful analysis. Moreover, the need for a proper introduction to the gaze-cued video recording was evident as the beginning of the test was spent talking about the gaze point rather than the test.

An important finding in testing phase one was the difficulty of having completely open tests in the system simulator. Previous research emphasizes the need to put the task in a realistic context to increase the validity of the data. Having a large test scope, where the thing being tested was only a small part of the overall task made the test difficult to control and resulted in each test being substantially different. Putting no limitations on participant interaction with the interface also increased the complexity of the analysis. Another finding was the limitation of only evaluating the reaction time for notifications. Because the test did not contain any objective related to finding usability issues with the feature, not much except how fast participants noticed the notification could be revealed. This does probably not reflect how eye-tracking would be applied in a real development process.

06

Testing Phase Two

This chapter accounts for the execution and results of tests C and D in testing phase two. Discussion about the findings and comparison of findings from the two tests is also included.

6.1. Execution

In this chapter, the test objectives and hypotheses of phase two are presented along with a description of the execution of the tests and analysis of results.

6.1.1. Test Objectives & Hypotheses

Testing phase one shed light on *level of interaction* and *test scope* as factors that needed further investigation. Moreover, previous research suggested eye-tracking is well complemented by thinking aloud and that auxiliary data should be applied to cross-validate eye-tracking results. Therefore, the aim of phase two was to see how limited versus unlimited interaction impacts eye-tracking data, its effectiveness in evaluating usability, and how other usability methods complement eye-tracking. The reasoning behind this relates to the way of working at Saab which often consists of validating specific features and windows in the interface. The objective was to determine the best way to design tests to facilitate this way of working. Adding on the two hypotheses from the previous phase, the following hypotheses were tested in phase two:

3. Eye-tracking data can more reliably be used to determine cognitive workload in tests with limited C2 interface interaction compared to tests with no limitations.
4. Eye-tracking metrics and visualizations can provide insights about the usability of the C2 interface that cannot be derived using traditional usability methods and metrics.
5. Gaze-cued retrospective think-aloud contributes with more usability insights in user tests with a wider scope compared to tests with a limited scope.

The goal was not only to be able to say something about how to use eye-tracking, but also how to use it in combination with other established usability methods. Therefore, it was decided that three different sources of data were to be used: eye-tracking metrics and visualizations, usability metrics, and gaze-cued retrospective think-aloud verbalizations.

6.1.2. Participants

Six participants were recruited for testing phase two, all of which were expert users and employed by Saab. The test included one woman and five men. One participant had undergone eye surgery. No participants had a current medical eye condition, other than impaired vision that could be compensated for with prescription glasses. In test C, one participant had used and been part of developing the feature that was being tested, two had used the feature before, and three had heard of but never used the feature.

6.1.3. Test Object, Test Design & Task

Testing phase two included tests C and D, performed by all participants and conducted back-to-back. Both tests were performed on the C2 system simulator, however with contrasting levels of permitted interaction. Similar to the previous phase, a within-subject test design was employed where all participants were exposed to all tasks. As described in phase one, each session began with the participant being asked to give informed consent to data collection and to fill out a questionnaire (see Appendix B). Some background information on the project, and the purpose being to test the technology and not user performance, was also presented. If needed, the eye-tracking glasses were then fitted with appropriate corrective lenses. A detailed description of the phase two test procedure can be found in Appendix C, a summary is presented below.

Test C

Test C was performed on a single feature in the C2 interface with known usability issues. The features consisted of two windows. The participant was limited to interacting with the feature alone and asked to not move any windows or open other windows than the ones specified. The tasks in test C were of low task scenario fidelity considering the limited and specified interaction. However, the depth of the tasks was higher as specific features were being tested.

Test C began with a description of the test followed by the calibration of the eye-tracking glasses, as described in phase one. This was followed by the starting of the recording and validation of the calibration, also presented in phase one. The participant was then given five tasks, one at a time, each involving accomplishing some specific assignment in the interface. After each task was completed, they were asked how certain they felt that they had completed the assignment and to rate their certainty on a scale from one to five, where one was “not at all certain” and five was “very certain”. Lastly, they were asked to comment on their rating.

Test D

Test D was a user test with a wider scope. The participants were allowed to use the entire C2 interface, with no limitations in interaction. Compared to test C, the task scenario fidelity was higher in test D as the task given was part of a scenario, mimicking the kinds of missions that operators perform. Nonetheless, no communication or other environmental factors contributing to the realism of the scenario or environment results in a lower testing environment fidelity and task scenario fidelity compared to the real-use context.

Test D, similar to test B, started with a mission brief provided by a test leader. The task was to identify hostile targets based on a number of criteria, this will be referred to as main tasks 1 and 2 (MT1 & MT2). MT1 and MT2 were described in a way that would force the participants to perform three other tasks in order to be able to complete the main tasks, these will be referred to as inferred tasks 1, 2, and 3 (IT1, IT2 & IT3). The aim was mainly to study their interaction with the three different windows contained in the three inferred tasks. The participants were given no limitations in their interaction with the interface and were allowed to solve the task freely. The test was terminated either when MT1 and MT2 had been completed, or when 15 minutes had passed.

Post-Test

After each test, the participants were asked to relax and look around the interface without interaction as a baseline measurement of their gaze was recorded. The recording was then ended, and the participants were asked to fill out a system usability scale form. Lastly, a gaze-cued retrospective think-aloud (GRTA) was performed with the recording of the just completed test. The GRTA was recorded using an audio recorder and video camera directed towards the screen to be able to connect verbalizations to the eye-tracking video recording that was viewed by the participant. The lessons from the GRTA in testing phase one led to participants receiving clearer instructions in testing phase two. During the GRTA, the participants were informed that it was not their performance that was being evaluated and asked to recall the thoughts they had while completing the assignments and to think aloud. When notable events occurred in the recording, they were asked to express their thoughts and reasoning at that moment. Besides that, questions were asked about why they were looking at a certain place based on their gaze point.

6.1.4. Analysis of Results

As previously mentioned, data from several sources was collected in testing phase two. These different data types were analyzed, and the results compiled separately. It was then possible to compare the overall usefulness of each data source.

Usability Metrics

The traditional usability metrics used in tests C and D were: number of clicks, number of errors, number of unnecessary actions, and task time. For test C, confidence was also measured, which was collected after each task was completed. To find the number of clicks for each respective task, the recording was studied with audio and each click was counted. Task completion time was also calculated for each task using the recording. In addition to this, the plain recordings (not containing gaze overlay) were studied to find errors and unnecessary actions. An error was defined as something that prevented or significantly hindered the participant from completing the task successfully. An unnecessary action, on the other hand, was defined as actions taken by the participant that did not take them closer to task completion, but at the same time did not affect whether the outcome was successful or not. Each error and unnecessary action were categorized based on the most common types for each respective test. The SUS scores were calculated in accordance with the method described in Chapter 3.6.1. All usability data was finally compiled for each test respectively and analyzed for insights.

Eye-Tracking Metrics & Visualizations

To allow for individual tasks to be analyzed separately and compared, TOIs had to be created in Tobii Pro Lab. For test C, the TOIs were created to span each task. For test D, the TOIs were created to span the time that the participant worked in the relevant window for the inferred tasks. Since the main tasks of test D spanned the whole test, no TOIs were created for these. A TOI was also created for each baseline measurement.

Similarly, to phase one, the following metric data was analyzed to determine the cognitive workload of the participants:

- fixation duration
- saccade amplitude
- peak saccade velocity
- pupil diameter

As in phase one, a decision was made to only consider long fixation durations as a high cognitive workload to simplify the analysis, as these seemed most applicable to C2 interfaces. Because of the known issues with pupil diameter measurements, that external factors affect the dilation more than cognitive workload, this data was first studied separately to determine its reliability. As the tests in this phase were conducted in a consistent environment, the data could be more reliable, although factors such as the participant looking off-screen remained. Based on this, the amount of weight given to the pupil diameter data could be adjusted. To compare the tasks based on cognitive workload, the mean value of each metric was calculated for each task and compiled for both tests.

A MATLAB script was created with the goal of finding segments in the data indicating a higher cognitive workload, and thus preventing the need to do this process manually. This script looked for instances when several of the metrics were either above (for fixation duration) or below (for saccade amplitude and peak saccade velocity) average. Since saccade velocity and amplitude are correlated, care was taken to not base a segment on solely these two. The script outputted intervals in the recording where these conditions were met, which was then manually investigated by referencing the recording. All intervals shorter than 3 seconds were removed in the filtering process due to lack of information, and to not base an interval on a single fixation and saccade. The gaze replay was studied at the identified intervals, noting what the participant was doing and any likely cause for an increased cognitive workload.

Heat maps were created for each task, with each heat map containing the data from as many participants as possible. Due to differences in window placement, it was not always possible to map the recording correctly on the still image, which is why some participants were left out in some cases. These heat maps were qualitatively analyzed based on what the task had been in each case and what settings were relevant.

Scan paths were also created for each participant and each task. These visualizations allowed for comparisons between participants, analyzing how successful and non-successful participants navigated with their gaze respectively, as well as looking for patterns that could reveal any additional usability problems.

Gaze-Cued Retrospective Think-Aloud

The audio recording that was collected during the GRTA in testing phase two was transferred into an intelligent verbatim transcript. The transcript was analyzed from two perspectives. Firstly, to find verbalizations regarding usability problems and the participant's mental model of the feature, as well as other comments regarding the interface. Secondly, to find verbalizations regarding how the method of gaze-cued retrospective think-aloud worked in practice, the participant's ability to recollect their thoughts during the test, and whether the gaze aided or interfered with the think-aloud.

Relevant quotes in the transcript were highlighted and categorized according to task and test participant. Subsequently, the problem areas were attributed to a specific usability problem type according to van den Haak (2008): terminology, data entry, comprehensiveness, feedback, relevance, formulation, visibility, completeness, graphic design, and structure. The number of unique usability problems verbalized in the GRTA was counted for tests C and D, to show how many usability problems could be found using the method. The number of usability problems expressed for each task in test C was counted as well. Here the problems were collected by counting the number of problems expressed by each participant in each task. Multiple statements by a participant referring to the same source of problem were counted only once and coded with a usability problem type. This was done to show what problems were most prevalent in each task, as these would likely be vocalized by multiple participants. Quotes related to how the method was experienced by the participant, how frequently they referred to the video or their gaze, whether they stayed in tune with the video, what adjustments were made to facilitate the method, and more, were also collected.

Studying Gaze Replays

The gaze replays were studied to find usability problems and further explain the cause of already found problems. This was done by playing the recording inside of Tobii Pro Lab and noting down times when potential usability problems could be identified. If the gaze also provided clues as to the reason for the problem, this was also noted down.

6.2. Results of Test C

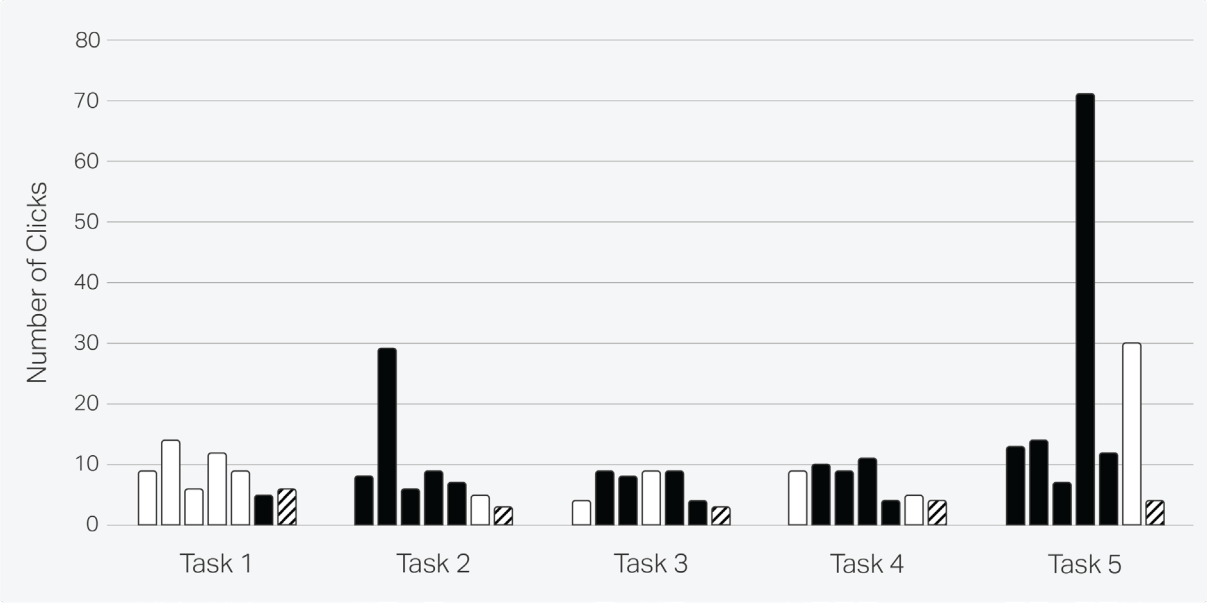
This chapter presents the results of test C in testing phase two. Included are the results of usability metrics and eye-tracking metrics, gaze-cued retrospective think-aloud, and an analysis of gaze replays. To conclude, a comparison of the results from each method employed in the test is presented.

6.2.1. Usability Metrics

The results showed that several tasks proved difficult for the participants, especially tasks 2-5 which most participants failed to complete correctly. All tasks had at least one participant completing it correctly, but most still used a substantially higher number of clicks than required (see Figure 8). Task time (see Figure 9), not independent from the number of clicks, shows a similar but more dispersed result.

Figure 8

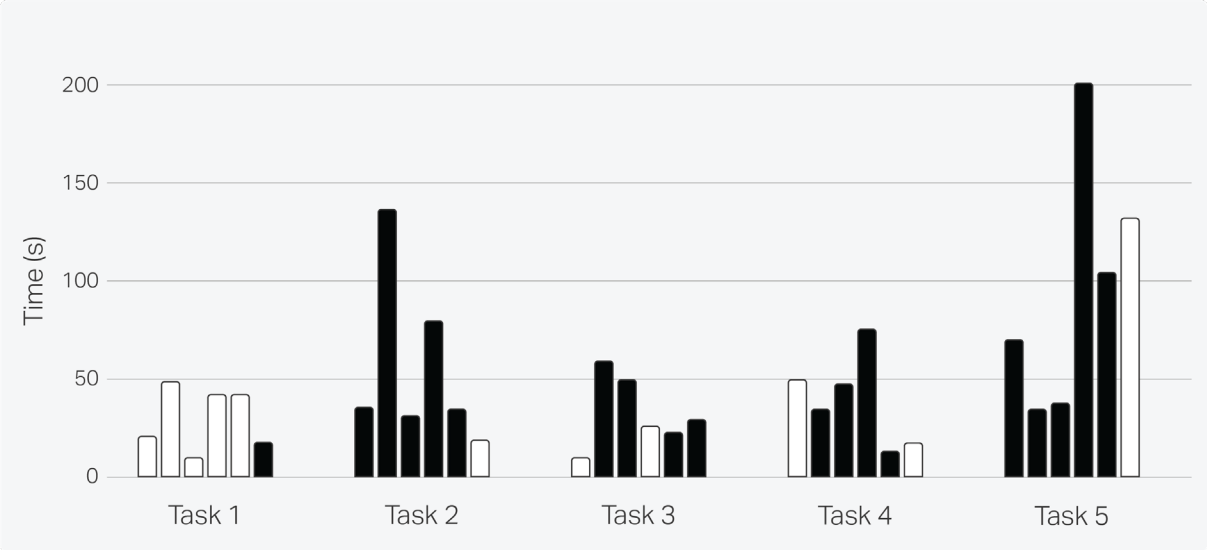
Number of Clicks in Test C



Note. Number of Clicks for Each Participant in Each Task of Test C. Color indicates whether the task was completed correctly or not (White = correct, black = not correct, striped = optimal). P6 in task 1 has a lower-than-optimal number of clicks due to a missed step.

Figure 9

Task Time in Test C

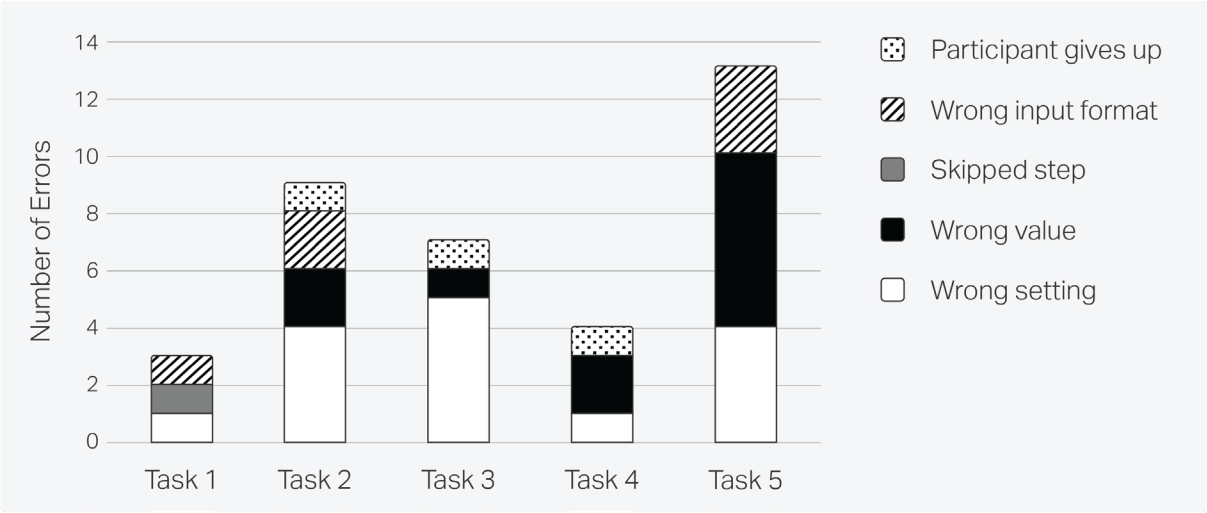


Note. The Task Time for Each Participant in Each Task of Test C. Color indicates whether the task was completed correctly or not (white = correct, black = not correct).

Errors were most prevalent in tasks 2 and 5 (see Figure 10), both involving similar settings. The most common error types were the participant choosing to change the wrong setting or changing the right setting to the wrong value. Unnecessary actions were most prevalent in tasks 1, 2, and 4 (see Figure 11). These point to an exploratory behavior was the participants look through the interface. The participants' confidence levels corresponded well to the success or failure of the task (see Figure 12), with only three confident answers (score 4-5) on failed tasks and three unconfident answers (score 1-2) on successful tasks.

Figure 10

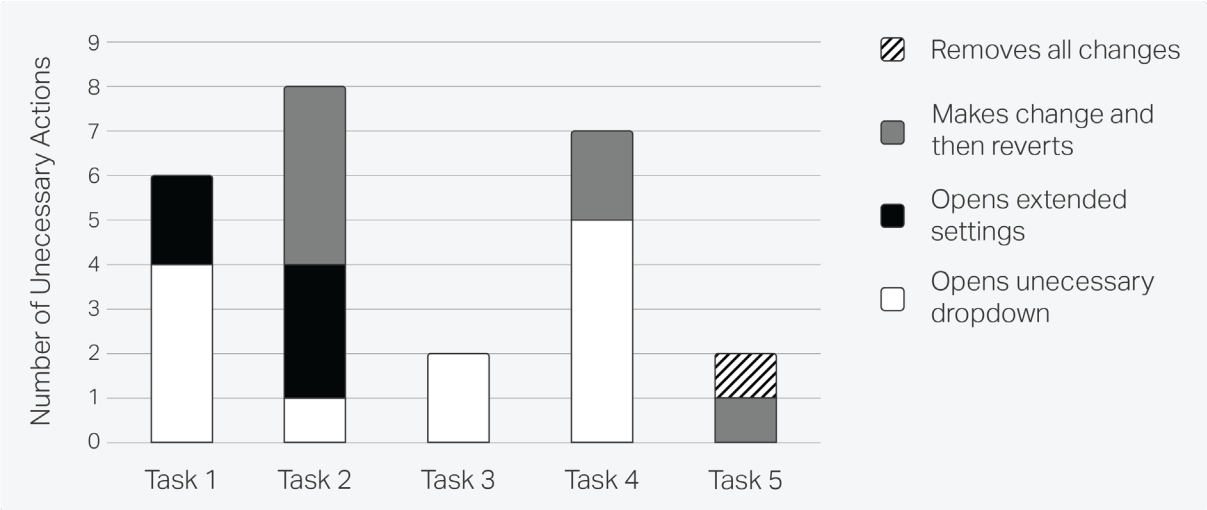
Errors in Test C



Note. The number of errors for all participants in test C, separated by task and categorized by error type.

Figure 11

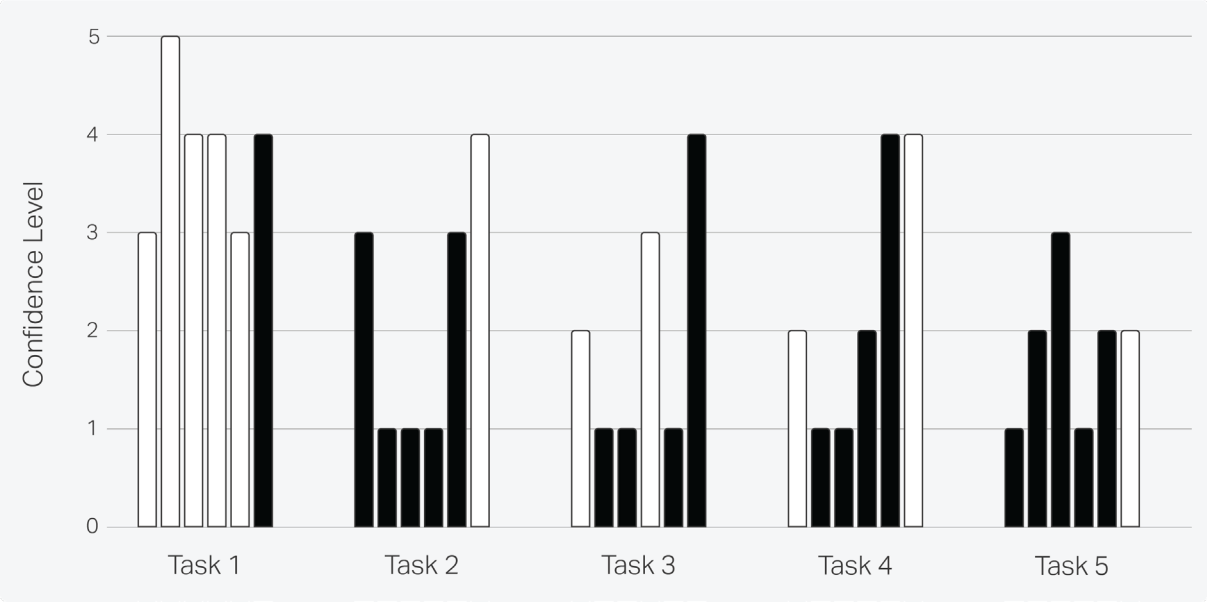
Unnecessary Actions in Test C



Note. The number of unnecessary actions for all participants in test C, separated by task and categorized by type.

Figure 12

Participants' Confidence Levels in Test C



Note. The participants' self-assessed confidence ranking on a scale from 1-5 for each task in test C. 5 means that the participant is very confident and 1 means they are not at all confident that they completed the task correctly. Color indicates whether a task was completed correctly or not (white = correct, black = not correct).

The results of the system usability scale (SUS) are presented in Table 5. The average score of 28.75 is to be considered poor. This shows that the subjective experience of the participants is that the specific feature that was tested is difficult to use, and has problematic usability.

Table 5

System Usability Scale Results for Test C

	P1	P2	P3	P4	P5	P6	Average
SUS score	32.5	0	35	27.5	40	37.5	28.8
Experience level	4	2	2	2	3	3	2.7

Note. The results of the system usability scale for test C (0-100) presented with the participants' self-valued experience level with the tested feature (1-5).

6.2.2. Eye-Tracking Metrics & Visualizations

For this test, the gaze sampling quality was between 96 and 99 percent for all recordings. The results from the metric data and eye-tracking visualizations will now be presented separately.

Metrics

Looking at the pupil diameter data separately, it became apparent that the largest increases in the diameter of the pupil mainly occur when the participant looks away from the screen. This was to be expected due to the lighting conditions of the testing environment. As a result, the pupil diameter cannot be used as a reliable measure to identify segments of higher cognitive workload in these tests. The data will however be presented along with the other metric data for later discussion.

Table 6 presents the collective averages for each metric and task, and the normalized data is presented in Table 7. The baseline measurement, which should have the lowest cognitive workload, indicates the lowest or a low cognitive workload for all metrics except for saccade amplitude. However, task 1, which should be the easiest, indicates a high cognitive workload for all metrics. Task 5, a harder task, at the same time, indicates a low cognitive workload. If each metric is given equal weight, each subsequent task would be considered to yield a lower cognitive workload than the previous.

Table 6

Cognitive Workload Metrics for Test C

	Task 1	Task 2	Task 3	Task 4	Task 5	Baseline
Fixation duration (ms)	960 (1275)	1118 (1235)	883 (1148)	766 (974)	878 (1269)	568 (682)
Pupil diameter (mm)	2.81 (0.14)	2.78 (0.16)	2.88 (0.18)	2.75 (0.13)	2.72 (0.11)	2.59 (0.11)
Peak saccade velocity (°/s)	207 (129)	258 (197)	272 (170)	235 (205)	286 (240)	281 (246)
Saccade amplitude (°)	7.13 (7.27)	6.85 (7.63)	7.05 (8.33)	6.13 (7.08)	12.14 (10.98)	7.23 (4.81)

Note. Data from test C showing eye-tracking metrics tied to cognitive workload separated by task, as well as the baseline measure. Each metric is accompanied by its mean value and standard deviation, presented in parentheses. The mean value reflects the average means among participants, while the standard deviation indicates the average standard deviation among participants within each category. The values represent averages for the entire task. Fixation duration and pupil diameter positively correlate with cognitive workload, while peak saccade velocity and saccade amplitude negatively correlate with cognitive workload.

Table 7

Normalized Cognitive Workload Metrics for Test C

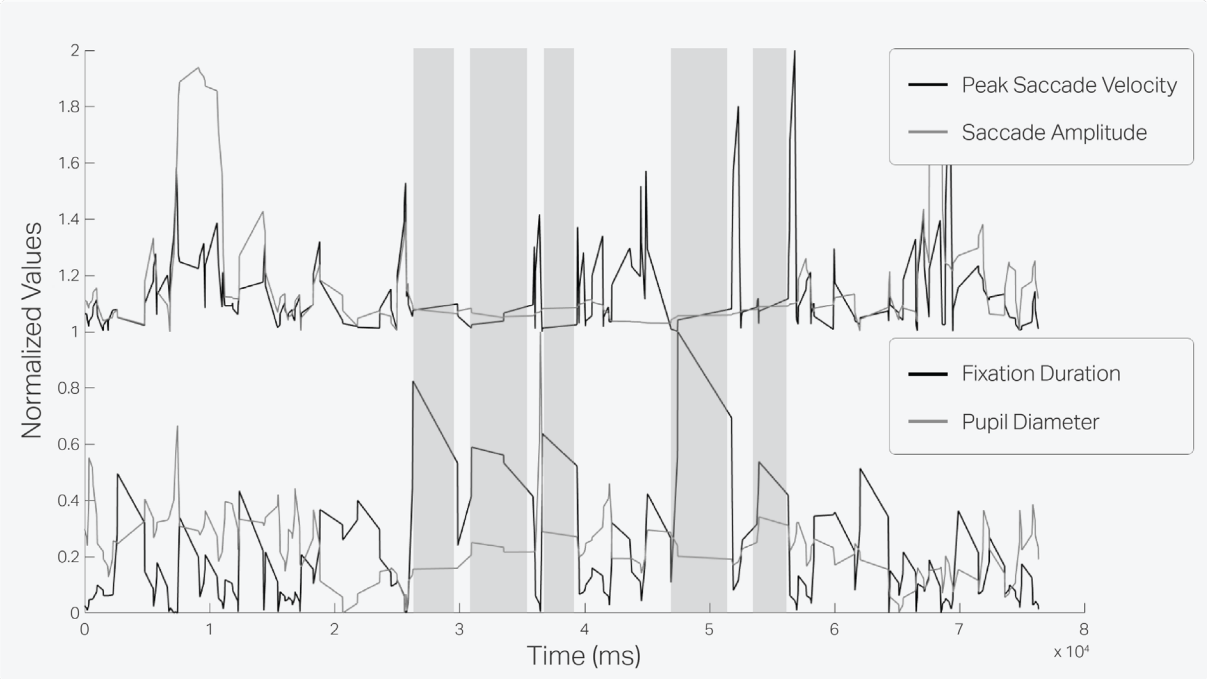
	Task 1	Task 2	Task 3	Task 4	Task 5	Baseline
Fixation duration	0.71 ↑	1.00 ↑	0.57 ○	0.36 ○	0.56 ○	0.00 ↓
Pupil diameter	0.78 ↑	0.67 ↑	1.00 ↑	0.57 ○	0.46 ○	0.00 ↓
Peak saccade velocity	0.00 ↑	0.65 ○	0.83 ↓	0.36 ○	1.00 ↓	0.93 ↓
Saccade amplitude	0.17 ↑	0.12 ↑	0.15 ↑	0.00 ↑	1.00 ↓	0.18 ↑

Note. Normalized values of the data presented in Table 6. Symbols indicate which third the value corresponds to, arrow up indicating higher cognitive workload, arrow down indicating lower cognitive workload and circle being in the middle. Fixation duration and pupil diameter positively correlate with cognitive workload, while peak saccade velocity and saccade amplitude negatively correlate with high cognitive workload.

In addition to comparing the entire tasks, the MATLAB script was used to identify specific moments during the task where the cognitive workload appears higher. Doing this for test C, consisting of shorter tasks (around 1 minute), proved unhelpful. With the intervals indicating high cognitive workload either being too short to gain any insights when using the raw data (see Figure 13) or covering most of the test when using a moving average (see Figure 14).

Figure 13

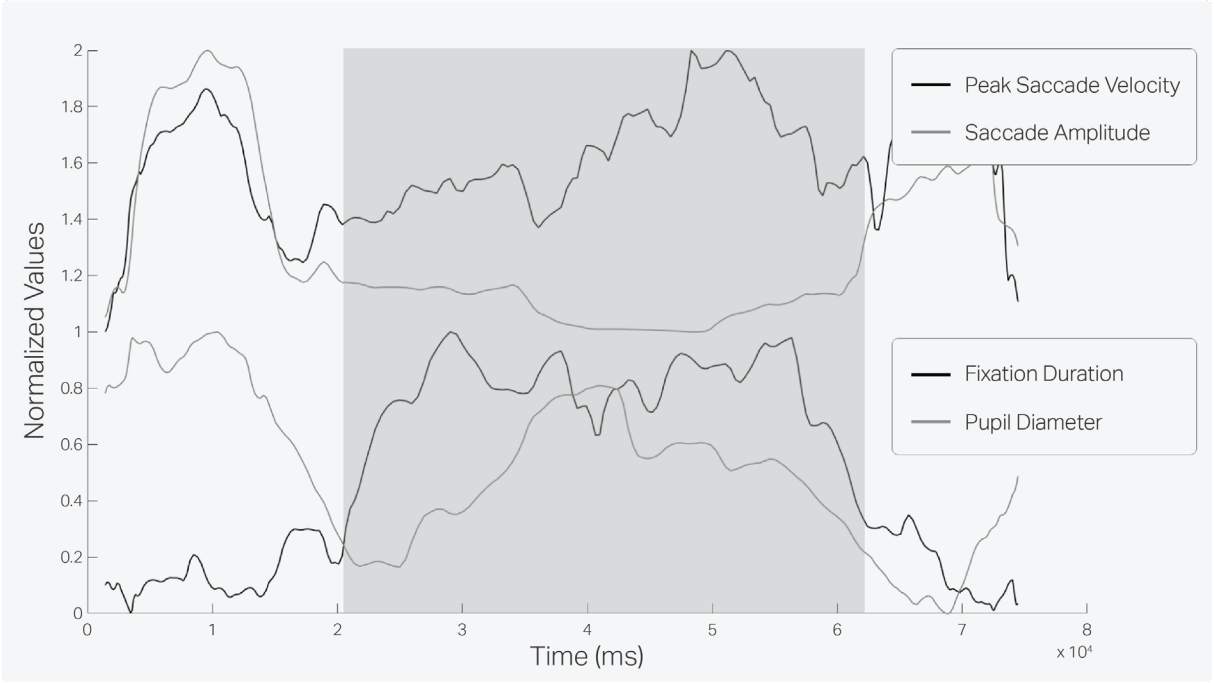
Example of Graph Without Moving Average



Note. Example showing a graph of the four cognitive workload metrics for a task in test C not using a moving average. Each metric is normalized separately as the values are different orders of magnitude. Peak saccade velocity and saccade amplitude have been offset by +1. High values for fixation duration and pupil diameter, as well as lower values for peak saccade velocity and saccade amplitude, indicate a higher cognitive workload. Several intervals indicating an increased cognitive workload are marked by the gray box.

Figure 14

Example of Graph with Moving Average



Note. Example showing a Graph of the four cognitive workload metrics for a task in test C using a moving average. Each metric is normalized separately as the values are different orders of magnitude. Peak saccade velocity and saccade amplitude have been offset by +1. High values for fixation duration and pupil diameter, as well as lower values for peak saccade velocity and saccade amplitude, indicate a higher cognitive workload. An interval indicating an increased cognitive workload is marked by the gray box.

Visualizations

The following heatmap visualizations provide an overview of which part of a window has received the most accumulated attention, from several participants, during each task. All images show a heat map overlay over a non-specific mockup of the tested window. All heat maps were generated using Tobii Pro Lab using the Tobii I-VT (Attention) filter, with a radius of 96 pixels.

Task 1 (see Figure 15, left): This heatmap shows a lot of activity on setting 9, which was not necessary for task completion. The participants seem to have looked at most settings even though only settings 1, 2, and 18 were required for task completion.

Task 2 (see Figure 15, right): This heatmap shows some activity around setting 8, which was the only required setting for this task. Most of the participants' attention does however seem to have been allocated to setting 7.

Figure 15

Heat Maps for Tasks 1 and 2

Note. The left image shows a heat map overlay for task 1 in test C. The generated heat map contains the aggregated data from five participants. The last participant's data was excluded due to the inability to perform assisted mapping due to different window placements. The right image shows a heat map overlay for task 2 in test C. The generated heat map contains the aggregated data from five participants. The last participant's data was excluded due to the inability to perform assisted mapping due to different window placements.

Task 3 (see Figure 16): These heatmaps show some activity around setting 11, which was the only required setting for this task. Participant attention was however very dispersed, showing that participants may have been unable to Figure out where to go.

Figure 16

Heat Maps for Task 3

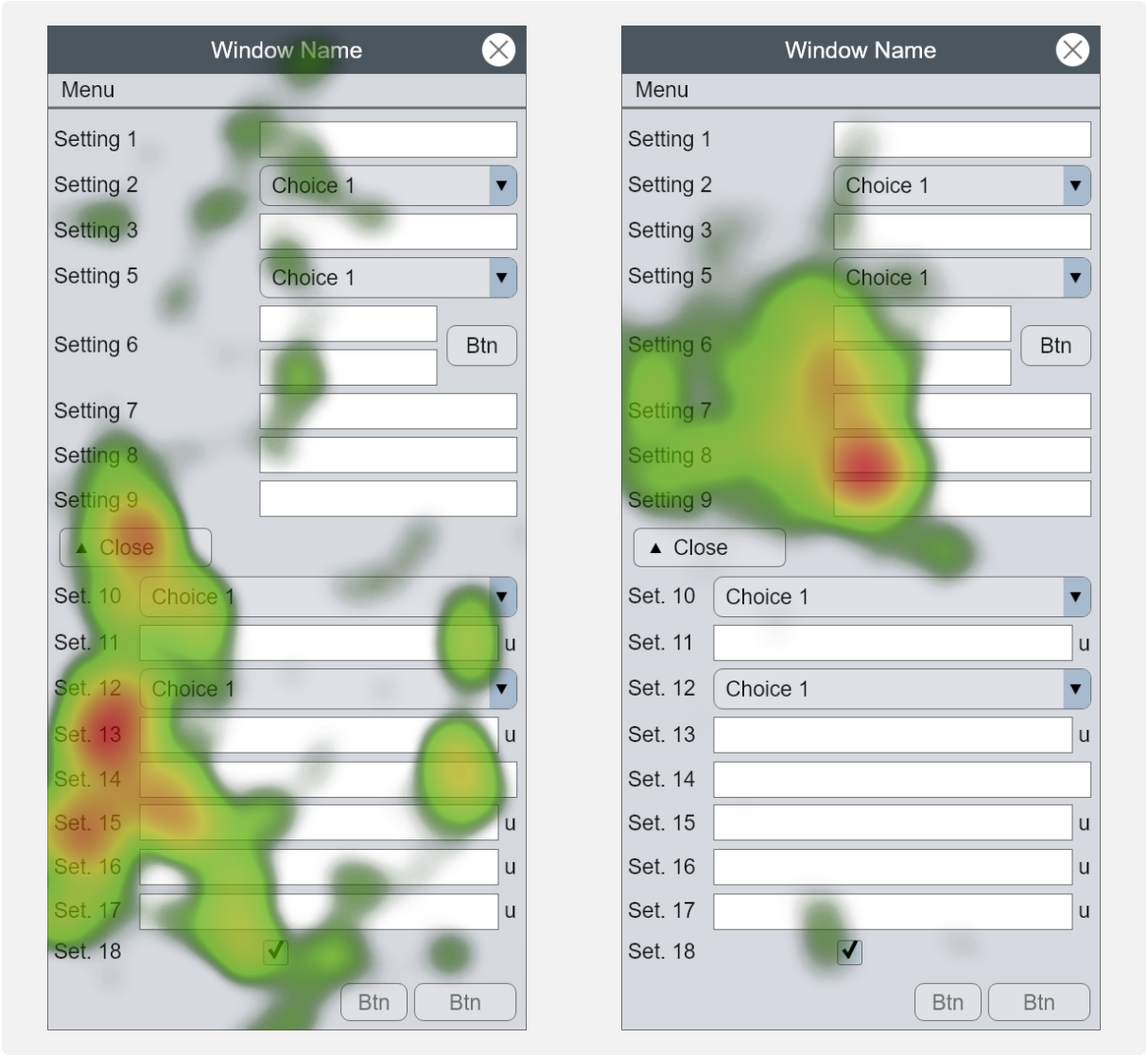
Note. This figure shows two heat map overlays for task 3 in test C. Due to different window placements, this task has been split into two different images. Each generated heat map contains the aggregated data from three participants.

Task 4 (see Figure 17, left): This heatmap shows some activity on settings 16-17, which were the only required settings for this task. However, the participants seem to have looked at setting 13-15 the most. The close button also seems to have gained a lot of attention.

Task 5 (see Figure 17, right): This heatmap shows the most activity on settings 6-9, with setting 8 gaining the most attention. For this task setting 7 and 8 was required. Why more time was spent on setting 8 cannot be discerned from the heatmap alone, but it could potentially indicate more uncertainty toward this setting.

Figure 17

Heat Maps for Tasks 4 and 5



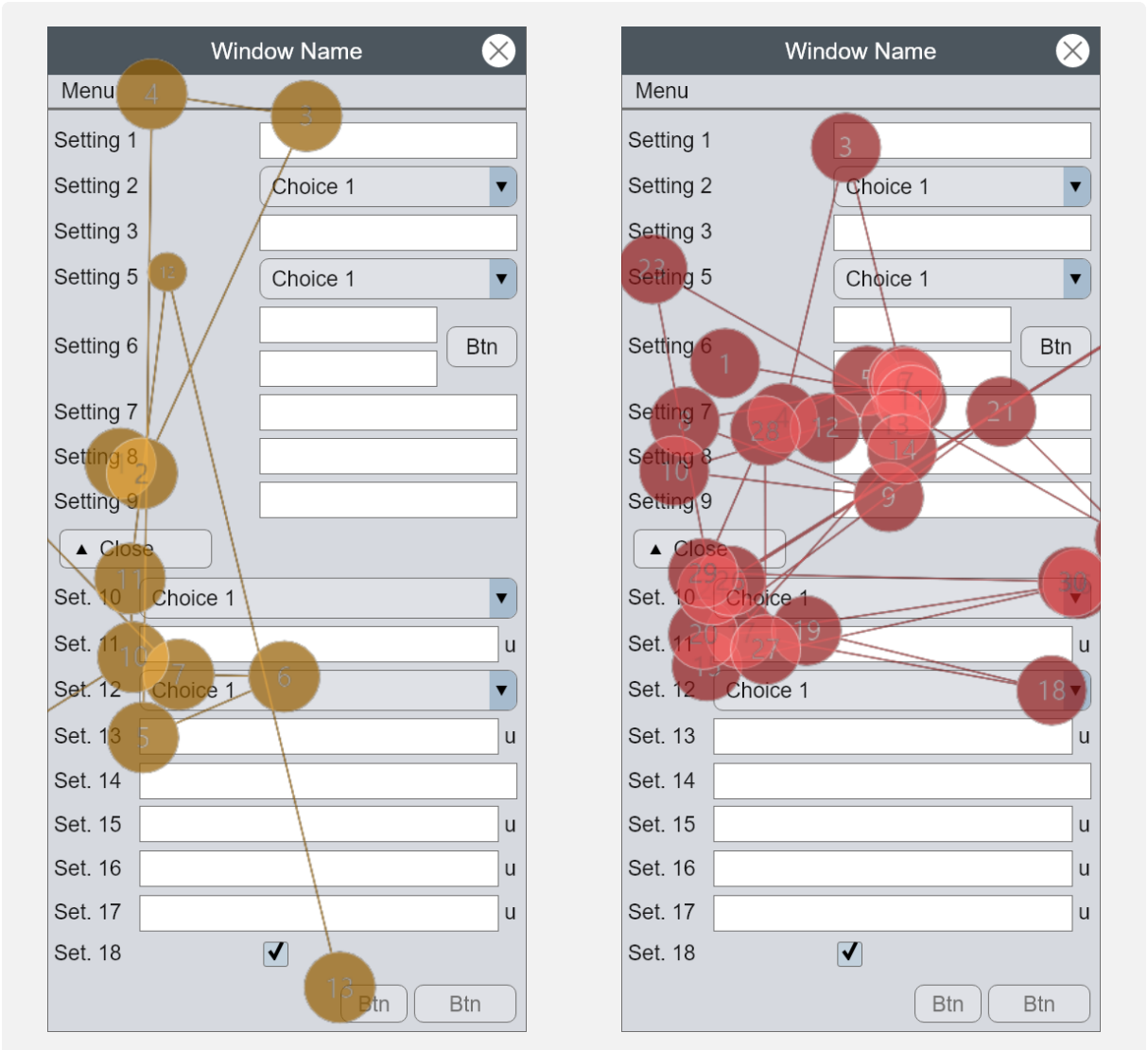
Note. The left image shows image shows a heat map overlay for task 4 in test C. The generated heat map contains the aggregated data from six participants. The last participant's data was excluded due to the inability to perform assisted mapping due to a different window placement. The right image shows a heat map overlay for task 5 in test C. The generated heat map contains the aggregated data from five participants. The last participant's data was excluded due to the inability to perform assisted mapping because of a software error.

From these heatmaps, it is possible to discern that the participants' attention was not always optimally distributed for the task. It is also possible to get a quick overview of which settings were given unnecessary attention, and also to what extent. The different task shows symptoms of different issues. The images for task 3 (Figure 16) for example show a dispersed searching behavior, indicating the participants not knowing what to look for, or finding what they are looking for. The image in task 5 (Figure 17, right) instead indicates that the participants know where the right settings are but are perhaps unsure of how to set their values.

Scan paths, dissimilarly from heat maps, only contain the data from one participant. Comparing a successful and unsuccessful attempt at task 3 (see Figure 18), some insights can be found. The example where the task was successfully completed shows the participant having fewer fixations, quickly deciding on the correct setting, and then changing the value. In the other example, where the task was not successfully completed, the participant instead looks extensively at two different settings, eventually deciding on changing the wrong one. The comparison reveals that some underlying usability issue causes participants to confuse the two settings.

Figure 18

Scan Paths Examples from Test C



Note. Two images showing scan paths for two different participants in task 3 of test C. Both images use a non-specific mockup of the tested window as background. The left image is from a participant who successfully completed the task (changed setting 11), and the right image is from a participant who filled out the wrong setting (setting 7).

6.2.3. Gaze-cued Retrospective Think-Aloud

Gaze-cued retrospective think-aloud resulted in insights concerning usability problems and participants' mental models. From the analysis of the transcribed verbalizations, several usability problem types were found which are presented in Table 8 along with examples.

Table 8

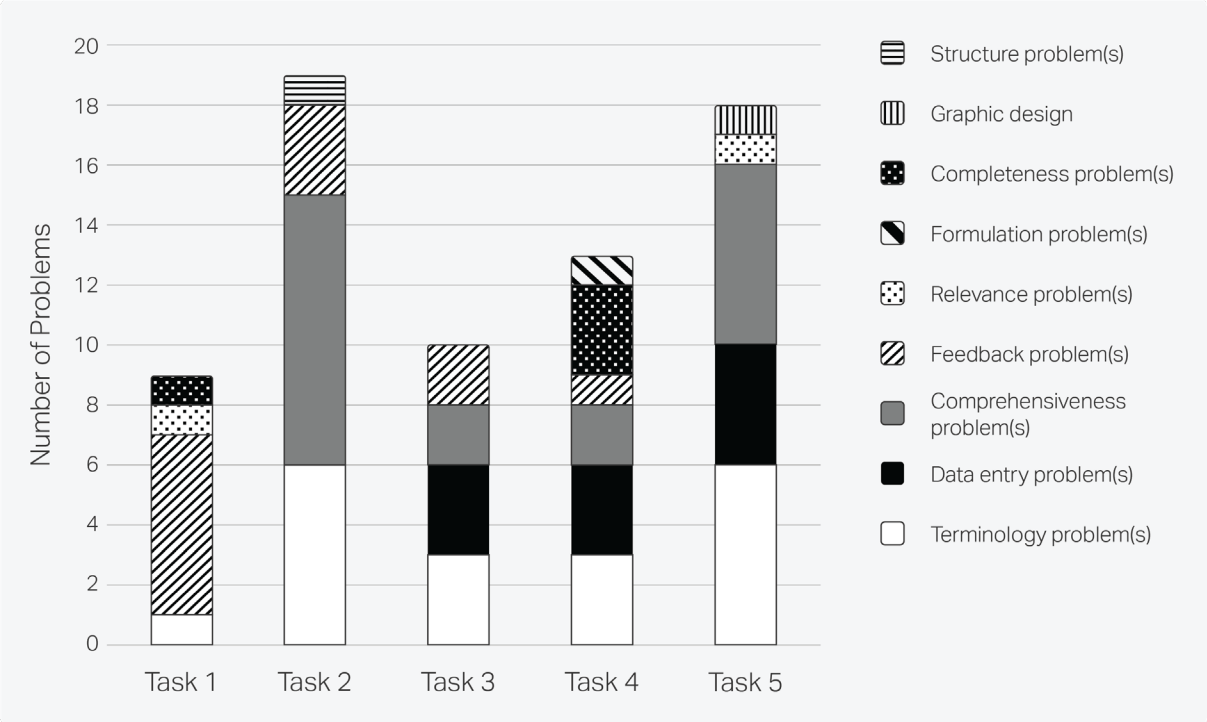
Usability Problem Types with Test C Examples

Terminology problem(s)	The participant does not understand the term for a certain setting.
Data entry problem(s)	The participant has trouble entering dates in the right format. The input format does not follow convention.
Comprehensiveness problem(s)	The participant does not understand what different settings do. The tooltip does not offer any additional help.
Feedback problem(s)	The feature fails to provide feedback on the impact that the adjustment of a setting has.
Relevance problem(s)	The participant feels that a certain setting should not be included.
Formulation problem(s)	The participant does not appreciate a particular formulation.
Visibility problem(s)	The participant fails to spot a particular element or information.
Completeness problem(s)	The participant feels that information is missing or more elaboration is needed.
Graphic design problem(s)	The participant does not appreciate the layout.
Structure problem(s)	The participant finds that the order of information is problematic.

In Figure 19 the number of problems in each category is presented for each task. The total number of unique usability problems that were expressed in test C was 37 (see Figure 20). It should be noted that although problems in Figure 20 are counted only once, many of the problems were mentioned by multiple participants as shown in Figure 19. Feedback was evidently an issue in task 1, a problem that is not included in other metrics as the main source of the problem was related to an element encountered after ending the task. Feedback was a recurring problem and many participants expressed that they did not know what effect their input would have and wished to see direct feedback. Many also assigned the wrong effects to certain settings. Terminology was undoubtedly a problem that, based on statements in the GRTA, was strongly related to the overall comprehension of the feature. As can be seen, comprehensiveness is a problem throughout and was often related to the mental model of the participant. The verbalizations reveal they did not comprehend how the settings were to be adjusted to achieve the goal they had, something that was amplified by the difficult terminology. Tasks two and five involved the same type of settings, hence similar problems with comprehensiveness. Data entry problems were often related to either unconventional input formats or the lack of guidance in what to enter in each field. The completeness problems that participants were experiencing in task 4 were due to missing default values in one setting.

Figure 19

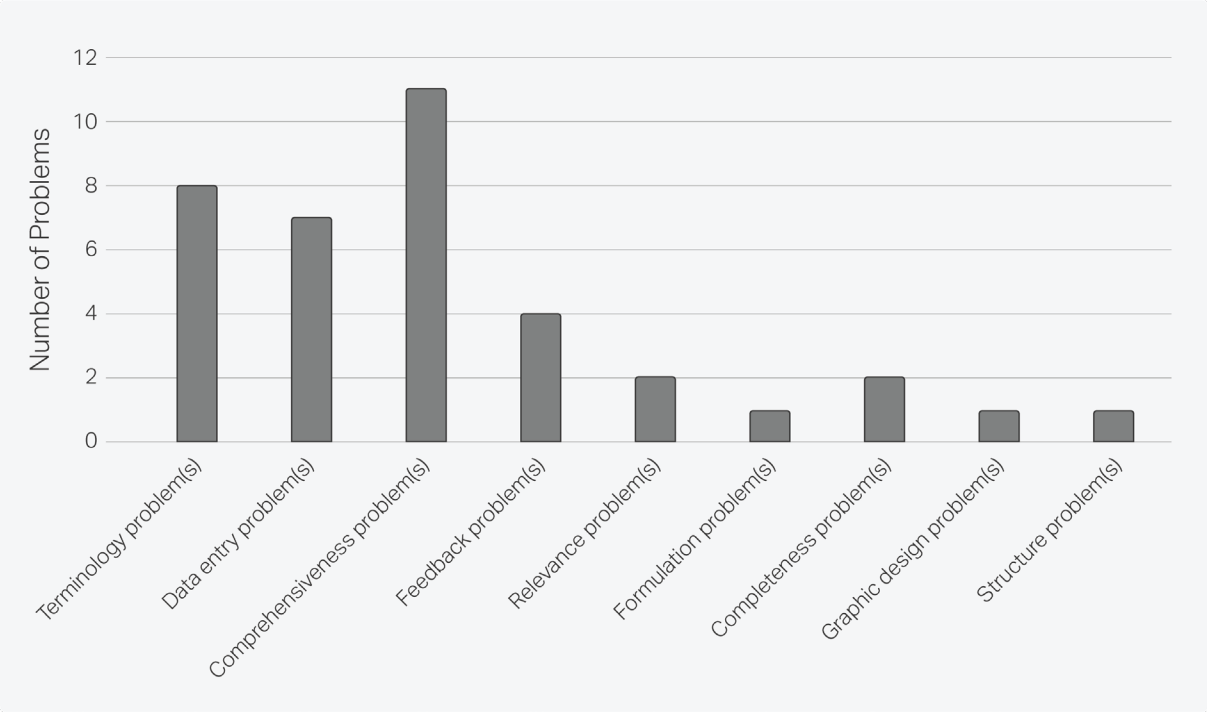
Usability Problems Expressed per Task in GRTA in Test C



Note. The number of expressed usability problems in the GRTA of test C, separated by task and categorized by type. Each participant can have multiple problems of the same type counted, although each specific problem will only be counted once for each participant.

Figure 20

Number of Unique Usability Problems Expressed in Test C



Note. The number of unique usability problems expressed in the GRTA of test C for each problem type. Each specific problem is counted only once, regardless of how many participants mentioned it.

After reviewing participants' statements, it is evident that there is a gulf of evaluation between the state of the interface and how well this is perceived and interpreted by the user. In the GRTA, participants expressed having trouble interpreting the interface and hence evaluating the correctness and effect of their input. The following quotes display this:

> "is this in a pre-setting mode or am I in real-time, am I doing it now, or am I doing it for later? Getting [feedback] there makes you think it's for later."

and

> "unsure whether the system is stuck waiting for something, am I supposed to do anything, or yeah, what's the next step?"

There is also a gulf of execution between the intentions of the user and what the system allows them to do. It became clear that the goal of the user was not facilitated by the interface. The quote below highlights this:

> "but I miss having, if you are going to set a [setting], then you have to have an [element] and an [element]".

This corresponds with the other results showing that participants had trouble completing the tasks successfully.

Besides this, a large number of other valuable insights were unveiled in the transcriptions, including participants' mental models, expectations, and emotions. Depending on what point of view the researcher is coming from, different aspects are of interest, but based on the result of this test it is clear that GRTA provides not only insight into specific problems but also serves as a window to a more holistic understanding of the human-machine interaction.

6.2.4. Analysis of Gaze Replays

Some examples found when analyzing the test C gaze replays include:

- A participant missed a step in one of the tasks. In the recording, it is obvious that the participant never looked where the setting in question was located, indicating that they did not see the setting, rather than seeing it and deciding to not change it.
- A participant chooses the wrong setting. In the recording, the participant's gaze can be seen switching between the right and wrong setting, finally opting for the wrong one. This indicates that choosing the wrong setting was not a case of missing it, but rather a lack of understanding leading to the wrong choice.

These types of examples, especially the second one, were commonly found when studying a gaze replay.

6.2.5. Comparing the Results

Test C showed a disconnect between the usability metrics and the eye-tracking metrics. The usability metrics indicate that tasks 2 and 5 were the most difficult, taking the most time and having the most errors. Task 1 had few errors and was quickly completed indicating that it was easier. The eye-tracking metrics on the other hand indicate that task 1 had the highest cognitive workload and task 5 the lowest. The cause of this is difficult to discern but could possibly be due to the later tasks being difficult but not demanding on the cognitive workload. The first task was possible for most people to solve, even with no prior experience of the feature. The later tasks on the other hand seemed to require previous knowledge and understanding. If the participants had the knowledge to complete the task, they did so with little effort. If they did not, they simply gave up and guessed the inputs. Therefore, it may not have been relevant to measure the cognitive workload in this test, as it does not necessarily reflect the usability level of the system. As task 1 by necessity was performed first by all participants, it is possible that they did in fact experience a higher cognitive workload during that task as they were familiarizing themselves with the feature. The difference in length between tasks 1 and 5 for most participants could also have influenced the mean values. Another possibility is that the eye-tracking data was influenced by some other unknown factor, and thus yielded results not aligned with the theory on the subject.

Analysis of gaze replays provided an additional perspective to the plain recording as it did offer some explanations for why errors occurred which would not have been possible without it. Heat maps provided an aggregated overview of the participants' attention. This adds additional insights into where problems in the interface lay by comparing it to where attention should be allocated. Scan paths also provide complementary information by allowing speculation on why a problem exists as shown by the example in the results.

The result of the GRTA aligns with that of the usability metrics with tasks 2 and 5 having the most expressed problems. The types of problems are not discernable from the usability metrics and only partly from the eye-tracking recording, thus displaying the value in allowing the participant to verbalize their general thoughts and provide reasoning as to why they were experiencing problems or performed errors. It also gives insight into their mental model which provides a holistic perspective of how they interpret the feature being tested.

6.3. Results of Test D

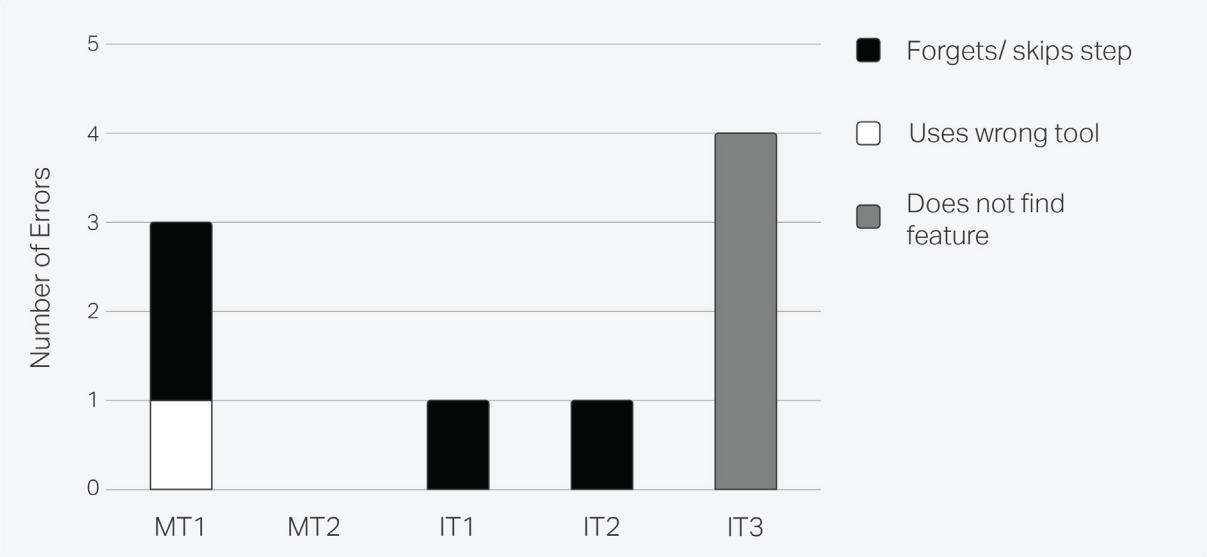
This chapter presents the results of test D in testing phase two. Included are the results of usability metrics and eye-tracking metrics, gaze-cued retrospective think-aloud, and an analysis of gaze replays. To conclude, a comparison of the results from each method employed in the test is presented.

6.3.1. Usability Metrics

The traditional usability metrics used in test D were: number of clicks, number of errors, number of unnecessary actions, and task time. Due to the more open and freer scenario of the main tasks, determining what was an error and unnecessary action proved difficult. Each participant had their own approach to solving the given tasks and thus performed widely different actions in different parts of the interface. While it is possible to determine the minimum viable actions required to complete the task and count everything else as unnecessary, this reflects a way of working that no operator would use in a real scenario and is therefore not a valid baseline. This is apparent in Figure 21 since few errors were discovered even though two participants failed to complete MT1 and four failed to complete MT2. The most common errors were skipping important steps, such as activating a feature, and not finding the correct place to perform IT3. What could be identified as unnecessary actions were exclusively repeat actions, where a participant for example turns something off and then on again right away, and actions that have no effect, such as configuring an unused feature (see Figure 22).

Figure 21

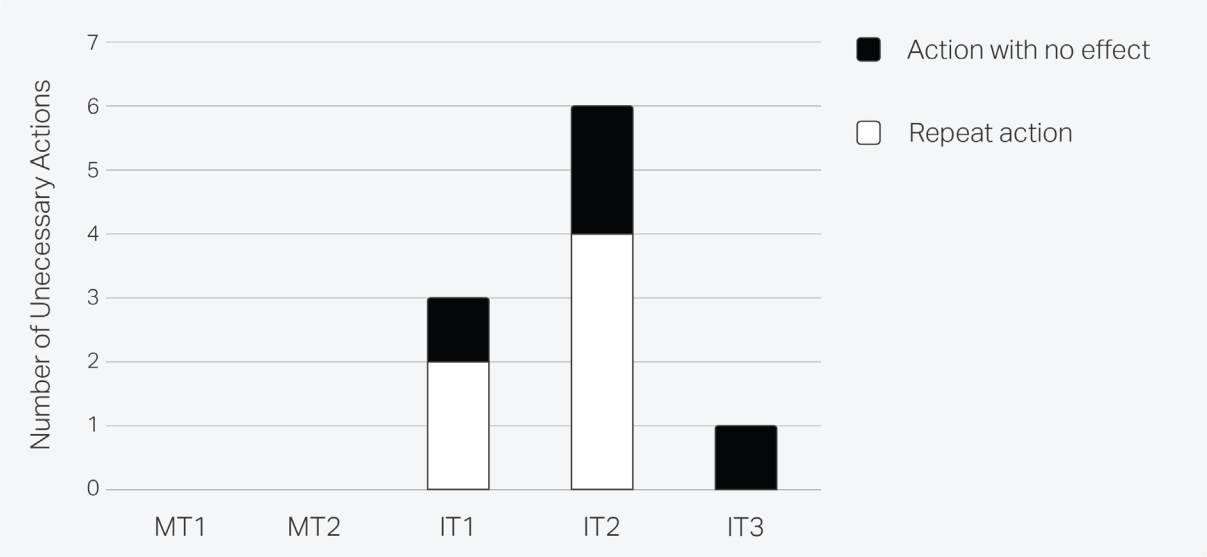
Errors in Test D



Note. The number of errors for all participants in test D, separated by task and categorized by error type.

Figure 22

Unnecessary Actions in Test D



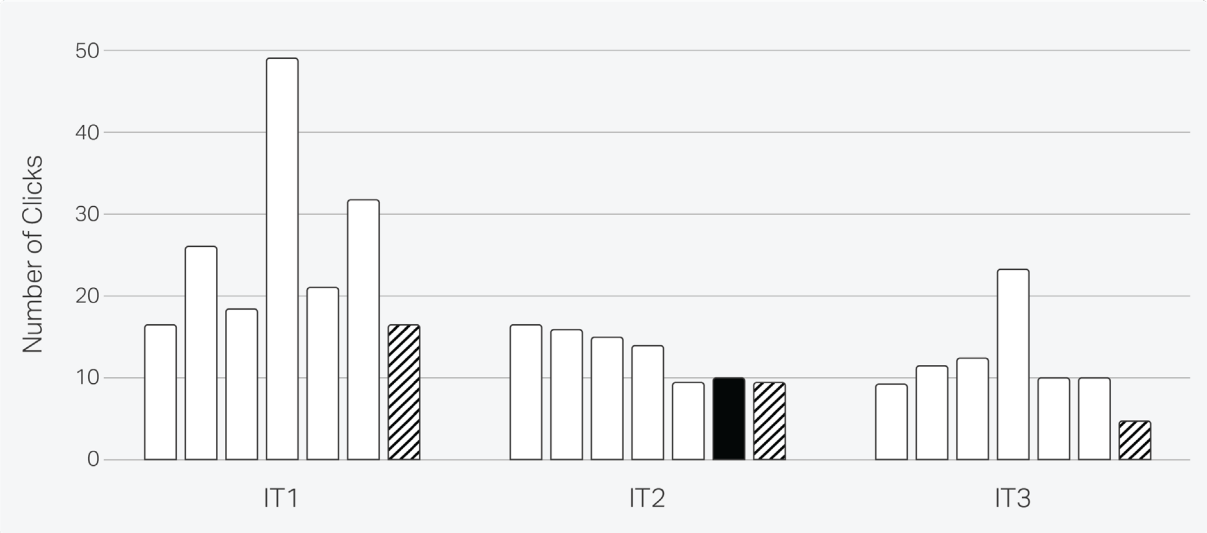
Note. The number of unnecessary actions for all participants in test D, separated by task and categorized by type.

Collecting the number of clicks for the main tasks is also not a viable measure since this does not necessarily reflect how easy it was for the participant to solve the task, or if they knew what they were doing or not. Since the inferred tasks are limited to a single window, the number of clicks could more easily be collected and compared to the optimal value, as that value is more apparent than in the main tasks. Figure 23 shows that participants had a consistent performance in IT2, close to the optimal value. Participants

had less consistency in IT1, with some participants performing far from optimally. IT3 showed the overall worst performance compared to the optimal, in several cases related to the participants not finding the correct feature, as presented in Figure 21. As task time (see Figure 24) and the number of clicks naturally correlate, these show similar results. However, since no optimal value exists for task time, this metric incorrectly (based on the number of clicks and errors) indicates IT3 as the easiest.

Figure 23

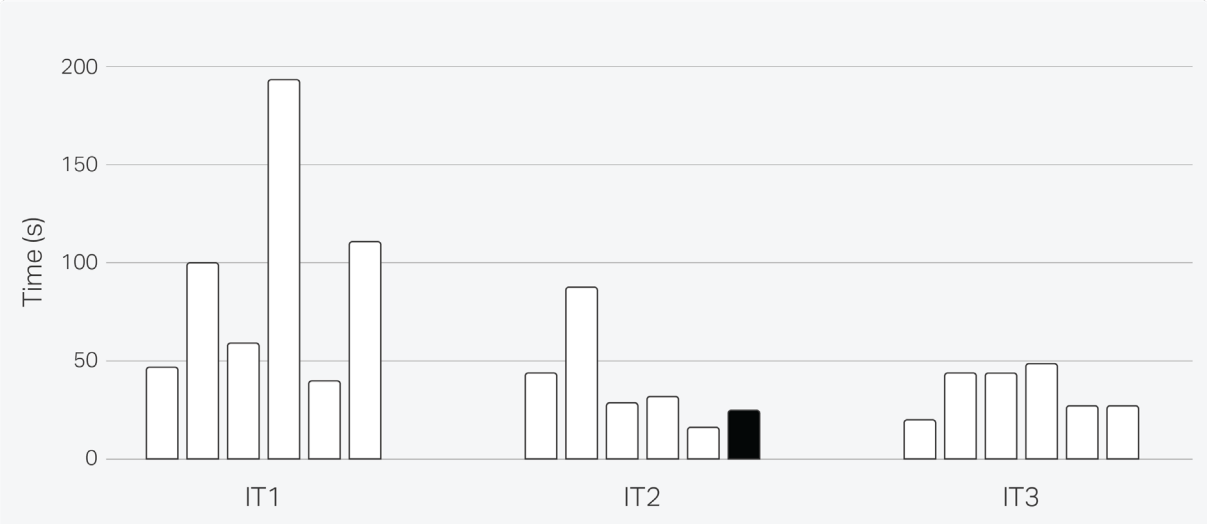
Number of Clicks in Test D



Note. The number of clicks for each participant in each task of test D. Color indicates whether the task was completed correctly or not (White = correct, black = not correct, striped = optimal).

Figure 24

Task Time in Test D



Note. The task time (in seconds) for each participant in each task of test D. Color indicates whether the task was completed correctly or not (white = correct, black = not correct).

The results of the system usability scale (SUS) are presented in Table 9. The average score of 65.4 is to be considered okay and is close to the general average for systems. This shows that the subjective experience of the participants is that the part of the system that was tested works okay as a whole. It should be noted that P4, who gave the lowest rating and is therefore pulling down the average, also rated their level of experience with the system the lowest.

Table 9

System Usability Scale Results for Test D

	P1	P2	P3	P4	P5	P6	Average
SUS score	45	80	77.5	37.5	75	77.5	65.4
Experience level	4	4	4	2	3	4	3.5

Note. The results of the system usability scale for test D (0-100) presented with the participants' self-valued experience level with the tested system (1-5).

6.3.2. Eye-Tracking Metrics & Visualizations

For this test, the gaze sampling quality was between 98 and 100 percent for all recordings. The results from the metric data and eye-tracking visualizations will now be presented separately.

Metrics

Table 10 presents the collective averages for each metric and inferred task, and the normalized data is presented in Table 11. Since the main tasks were continuous, they were not possible to isolate in the recording and are therefore not presented. The baseline measurement, which should have the lowest cognitive workload, indicates the lowest or a low cognitive workload for all metrics but the saccade amplitude. Which task out of IT1 and IT2 had the highest cognitive workload is difficult to determine definitively, but IT3 at least seems to yield a somewhat lower workload. Overall, there is not sufficient consistency in the data to draw any major conclusions from it.

Table 10

Cognitive Workload Metrics for Test D

	IT1	IT2	IT3	Baseline
Fixation duration (ms)	719 (1007)	750 (879)	634 (643)	587 (523)
Pupil diameter (mm)	2.81 (0.11)	2.84 (0.12)	2.80 (0.13)	2.76 (0.11)
Peak saccade velocity (°/s)	247 (188)	249 (124)	242 (168)	273 (223)
Saccade amplitude (°)	6.18 (4.37)	6.94 (3.92)	6.38 (4.72)	6.59 (4.16)

Note. Data from test D showing eye-tracking metrics tied to cognitive workload separated by task, as well as the baseline measure. Each metric is accompanied by its mean value and standard deviation, presented in parentheses. The mean value reflects the average means among participants, while the standard deviation indicates the average standard deviation among participants within each category. The values represent averages for the entire task. Fixation duration and pupil diameter positively correlate with cognitive workload, while peak saccade velocity and saccade amplitude negatively correlate with cognitive workload.

Table 11

Normalized Cognitive Workload Metrics for Test D

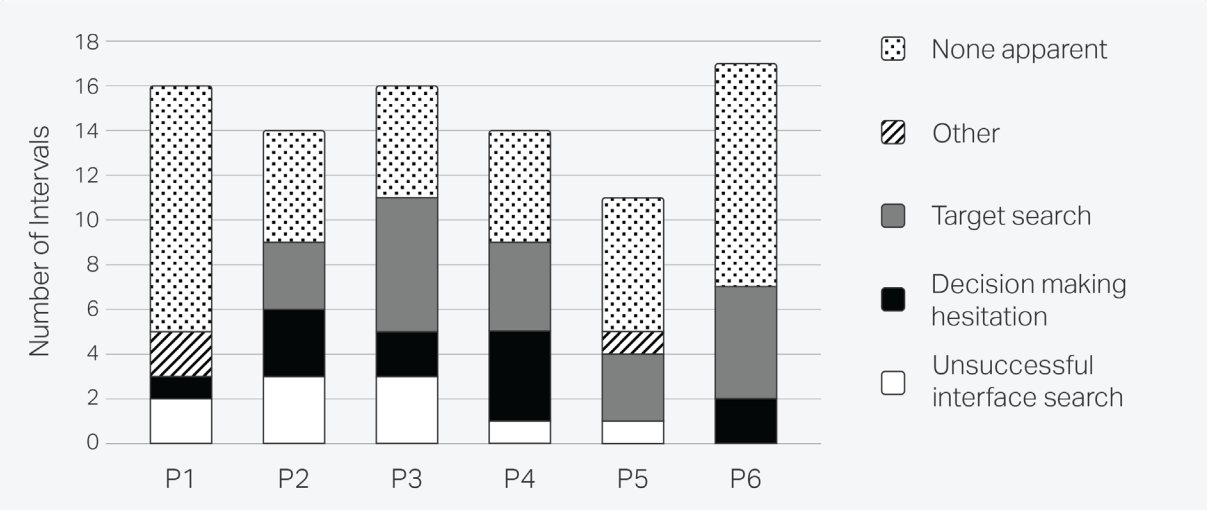
	IT1	IT2	IT3	Baseline
Fixation duration	0.82 ↑	1.00 ↑	0.29 ↓	0.00 ↓
Pupil diameter	0.68 ↑	1.00 ↑	0.44 ○	0.00 ↓
Peak saccade velocity	0.15 ↑	0.23 ↑	0.00 ↑	1.00 ↓
Saccade amplitude	0.00 ↑	1.00 ↓	0.25 ↑	0.54 ○

Note. Normalized values of the data presented in Table 10. Symbols indicate which third the value corresponds to, arrow up indicating higher cognitive workload, arrow down indicating lower cognitive workload and circle being in the middle. Fixation duration and pupil diameter positively correlate with cognitive workload, while peak saccade velocity and saccade amplitude negatively correlate with high cognitive workload.

The MATLAB script identified 11 to 17 segments in each 9-to-15-minute recording possibly containing a higher cognitive workload. Each such segment was from 3 to 70 seconds. Many segments contained no apparent reason for an increased cognitive workload, but several contained moments when the participant was either looking for something in the interface that they could not find, hesitating when faced with options, or looking for targets on the map (see Figure 25). Some features and windows were represented to a higher extent among participants in these segments, indicating that these are either subject to issues or that they are more burdensome by nature.

Figure 25

Possible Workload Interval Cause

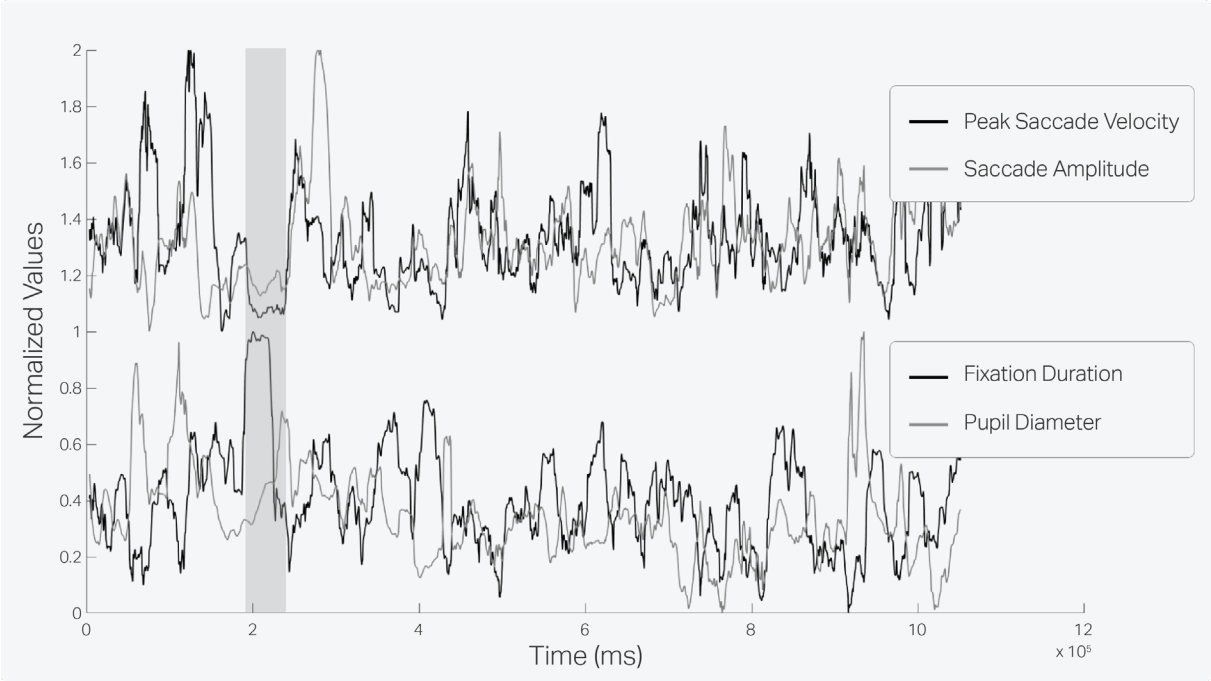


Note. The number of segments indicating a high cognitive workload for each participant, categorized by the observed possible cause for the cognitive state. None means that no apparent reason for an increased cognitive workload could be observed, not that no such increase existed.

The threshold for what should count as a segment was purposely set low, to avoid missing any relevant data, which is possibly why several segments contained no obvious cause of high cognitive workload. In three of the recordings, however, one segment stood out as having the highest fixation durations and some of the shortest saccade velocities and saccade amplitudes of the recording. The first example starts when the participant is faced with a problem where its cause seems uncertain (see Figure 26). During the segment, the participant tries to find the cause and solve the problem. The segment then ends right as the problem is solved. During all D-tests, this was the clearest example of a participant having to problem-solve their way to a solution.

Figure 26

Graph of Cognitive Workload Metrics – Example 1

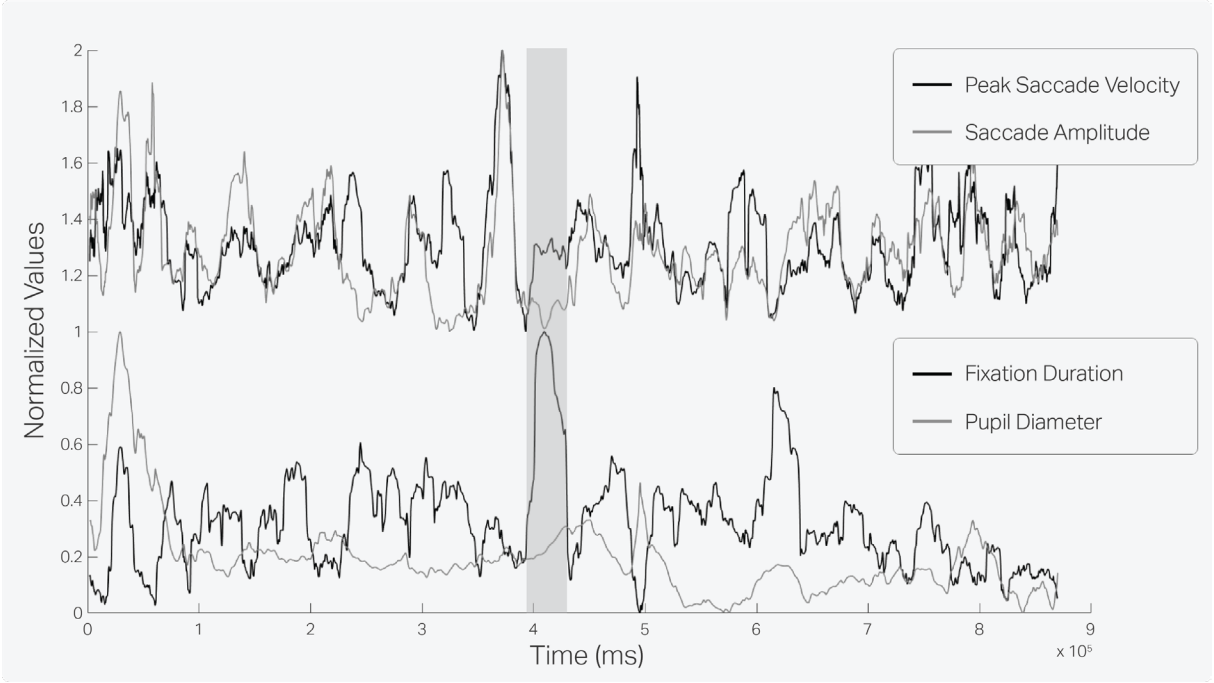


Note. A graph of the four cognitive workload metrics from one participant, for the entirety of test D, using a moving average. Each metric is normalized separately as the values are different orders of magnitude. Peak saccade velocity and saccade amplitude have been offset by +1. High values for fixation duration and pupil diameter, as well as lower values for peak saccade velocity and saccade amplitude, indicate a higher cognitive workload. An interval indicating an increased cognitive workload is marked by the gray box.

The second example occurred when a participant was looking for a setting without success (see Figure 27). In the recording, the participant can be seen scratching their head, possibly confirming a high cognitive workload at this moment. Another interesting aspect is the increase in pupil diameter from 3.01 mm to 3.40 mm during the segment, even though the participant remained still, only looking at the same part of the screen and lighting conditions remaining unchanged. This could indicate that the increase in pupil diameter is due to an increased cognitive workload and that while pupil diameter could not be used to locate the interval, it could be used to confirm it. A similar increase in pupil diameter can also be seen in the interval in Figure 26.

Figure 27

Graph of Cognitive Workload Metrics – Example 2

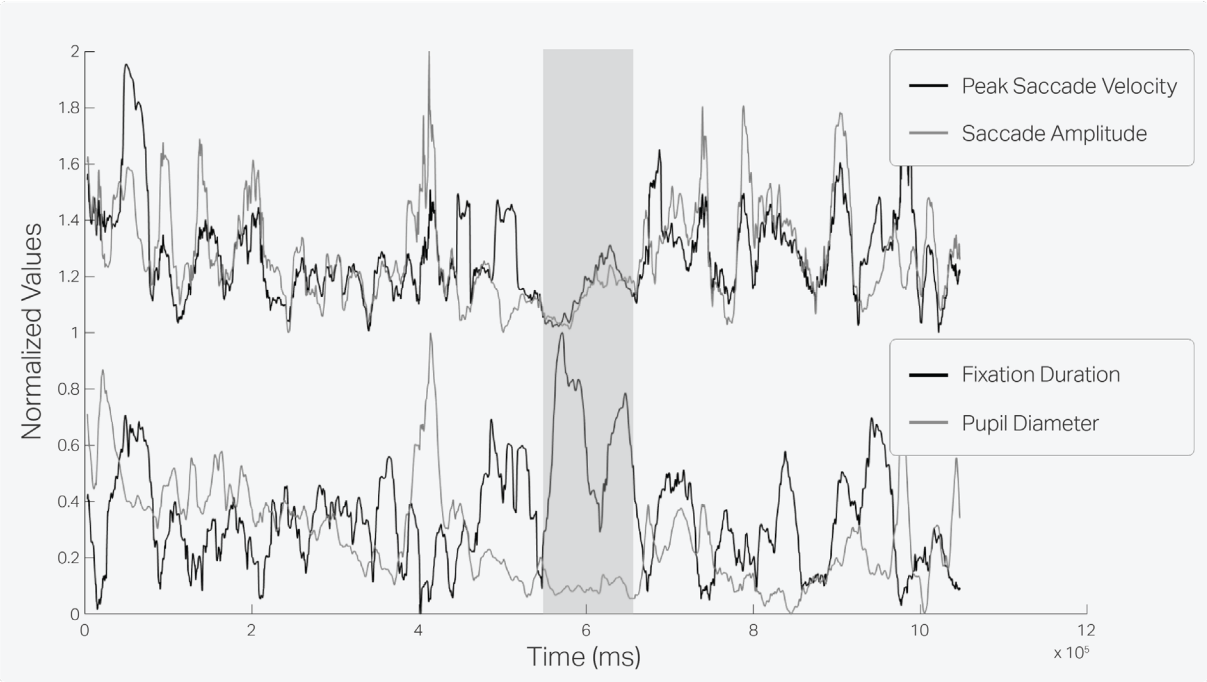


Note. A plot of the four cognitive workload metrics from one participant, for the entirety of test D, using a moving average. Each metric is normalized separately as the values are different orders of magnitude. Peak saccade velocity and saccade amplitude have been offset by +1. High values for fixation duration and pupil diameter, as well as lower values for peak saccade velocity and saccade amplitude, indicate a higher cognitive workload. An interval indicating an increased cognitive workload is marked by the gray box.

The third example takes place as the participant carries out part of the main task (see Figure 28). In the recording, the participant can be interpreted as hesitating or carefully performing each action to not make any mistakes.

Figure 28

Graph of Cognitive Workload Metrics – Example 3



Note. A plot of the four cognitive workload metrics from one participant, for the entirety of test D, using a moving average. Each metric is normalized separately as the values are different orders of magnitude. Peak saccade velocity and saccade amplitude have been offset by +1. High values for fixation duration and pupil diameter, as well as lower values for peak saccade velocity and saccade amplitude, indicate a higher cognitive workload. An interval indicating an increased cognitive workload is marked by the gray box.

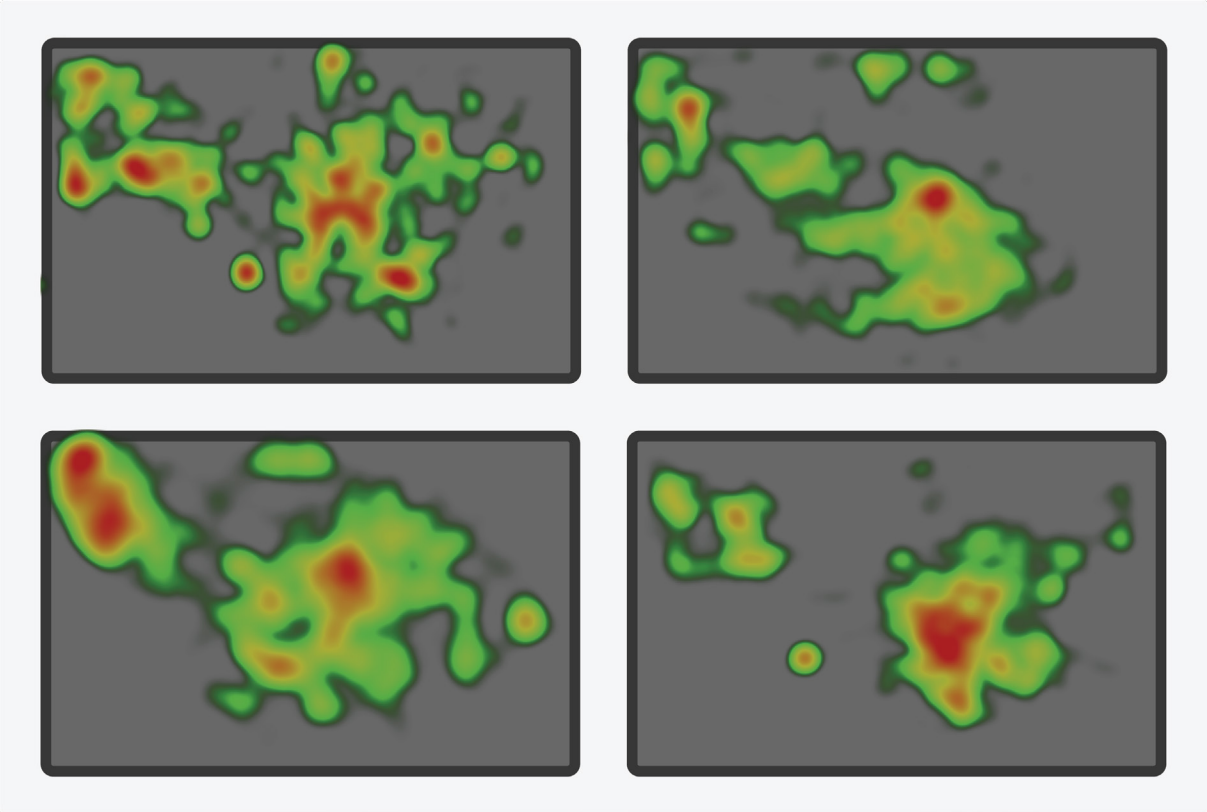
The issue with analyzing these results lay in that it is difficult to determine whether any of the identified segments actually corresponded with a high cognitive workload, objectively or as subjectively experienced by the participant. It is possible that the three most significant examples (presented in Figure 26-28) were the only ones to actually have a measurable effect. It is also possible no measurable cognitive workload was present at all. Some differences in the metrics can also surely be attributed to what is being looked at. Observing points on a map will for example naturally have longer saccade amplitudes than reading. What is clear, however, is that certain actions and features tend to appear more frequently in these segments than others, allowing this data to be used to learn about what can cause an increased cognitive workload, and provide a basis for future improvements.

Visualizations

The heatmap visualizations in Figure 29 provide an overview of which areas of the screen have received the most accumulated attention during the entirety of test D for each participant. All heat maps were generated using Tobii Pro Lab using the Tobii I-VT (Attention) filter, with a radius of 96 pixels. Since the heat maps are mapped onto a still image (which has been removed in Figure 29), the heat maps provide no insights into what features are being used during the tests. The functionality of the eye-tracking software is not adapted to the dynamic nature of the interface. As such, it cannot be used to identify usability problems or other issues in this case. However, if the goal for example is to study how operators utilize and manage screen space during a session, heat maps created like this could provide valuable insight.

Figure 29

Heat Maps from Test D



Note. Examples of four different participants' attention distribution during the entirety of test D, presented as heat maps.

6.3.3. Gaze-cued Retrospective Think-Aloud

The majority of verbalizations from this GRTA in test D were reports of the participant's actions rather than reasoning and thoughts. The number of expressed unique usability problems were 18 in total. The most commonly expressed problem in Test D was related to visibility (see Figure 30). Since all operators performed the task in different ways, all faced different issues related to visibility and not finding what they were looking for, such as:

> "Now I am searching for [feature], because I can't find it."

or

> "I find these complicated, so that you know that you are connecting the right things. At one point I missed highlighting it and therefore it did not connect right away, and that should probably be clearer"

The second most common is feedback which was often related to not receiving visible feedback on whether something is active or not.

> "I was trying to get feedback by clicking it, but don't think anything happened."

and

> "When you see yellow, it is not certain that you have everything activated."

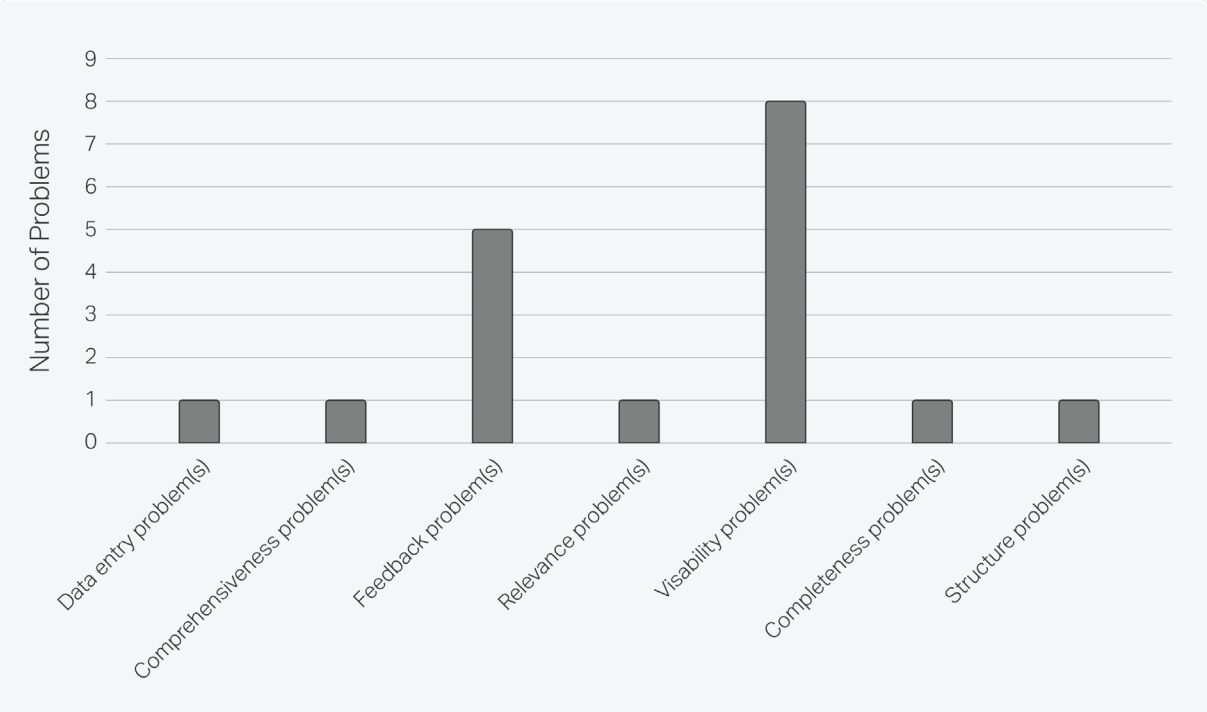
Comprehensiveness problems were also mentioned and often related to unexpectedly not receiving the information the participant was seeking and not understanding what type of setting has been put in place.

> "Later on, I didn't really understand what kind of setup we had in this scenario, what [feature] was selected and someone made settings there that I didn't really understand."

The GRTA unveiled insights besides those that can be directly attributed to a specific usability problem. For example, the task was performed in five different ways, with almost all participants performing it in their unique way.

Figure 30

Number of Unique Usability Problems Expressed in Test D



Note. The number of unique usability problems expressed in the GRTA of test D for each problem type. Each specific problem is counted only once, regardless of how many participants mentioned it.

6.3.4. Analysis of Gaze Replays

The main insights gained from studying the gaze replays are related to where and how the participant moves their gaze at different moments. Some examples found when analyzing test D include:

- A participant missed a step in one of the inferred tasks. In the recording, it is obvious that the participant never looked in the area where the button in question was located, indicating that they did not see the button, rather than seeing it and not pressing it.
- A participant missed a step in one of the inferred tasks. In the recording, the participant's gaze can be seen passing over the necessary setting but ignoring it. This indicates that they did see the setting but found no reason to use it due to a lack of knowledge or a lapse in memory.
- A participant opens the correct window for the task but closes it again without action. In the recording, the participant can be seen navigating between tabs in a window, not knowing which one contains the right setting. The participant opens the right tab, but their gaze never inspects its contents and therefore continues with the search. This indicates that the problem lies with the terminology in the tabs, and not how the setting is presented as a whole.

- A participant hovers with the mouse over a window's close button and hesitates. In the recording we can determine that the participant hesitates due to not yet receiving feedback from the interface, actively searching for it. This indicates that there is no problem with the button and that the participant only wanted to make sure that the goal was achieved before closing the window.

These types of examples, either indicating problems or dismissing them, were commonly found when studying a gaze replay.

6.3.5. Comparing the results

Determining what was an error or an unnecessary action was difficult due to the freedom of interaction that was provided in test D. Although finding intervals in the eye-tracking data that indicated cognitive workload did provide a place to look for issues, the current threshold provided a too large part of the recording to be of any real assistance. Instead, only the largest peaks in the data proved to be of substantial interest.

As with test C, it is difficult to see any clear connection between the normal usability metrics and the cognitive workload. This is made more difficult because of the difference within the usability metrics, with IT3 having the most errors but a low task time and number of clicks. Also, the difference within the eye-tracking metrics, with different metrics indicating high and low load within the same test, makes conclusions difficult. The cause of this discrepancy can be many things, but the participants' ability to interact freely with the interface is one possible cause.

Even though many of the usability problems could have been detected through a normal screen recording, it would in that case be impossible to determine whether the participant for example was unsure what to choose from a dropdown menu or if they were simply looking somewhere else at that moment. The qualitative analysis of the gaze replay did provide this perspective, deepening the usability insights. In addition, the verbalizations from GRTA provided a unique source for their reasoning in those moments. Many problems that were expressed in the GRTA were not found through any other methods, for example, the participants' expectations and intentions.

07

Review of Test Design & Evaluation Methods

This chapter discusses the test design factors and evaluation methods that were tested in this study. The hypotheses that were posed at the beginning of each phase will also be evaluated based on the findings presented in Chapters 5 and 6. The hypotheses were:

1. Eye-tracking data from a low-fidelity system prototype is comparable to data from a high-fidelity system prototype of a C2 interface.
2. Non-expert eye-tracking data can be used as a substitute for expert data.
3. Eye-tracking data can more reliably be used to determine cognitive workload in tests with limited C2 interface interaction compared to tests with no limitations.
4. Eye-tracking metrics and visualizations can provide insights about the usability of the C2 interface that cannot be derived using traditional usability methods and metrics.
5. Gaze-cued retrospective think-aloud contributes with more usability insights in user tests with wider a scope compared to tests with a limited scope.

7.1. Test Design

It was clear from both testing phases that the type of test (exploratory, comparative, assessment, or verification) being performed affected which evaluation methods could be used. This in turn affected other factors such as test time, participant characteristics, and scenario. Choosing the right test type, according to the goals and needs of the study, is therefore essential. Exploratory tests seem effective when wanting to learn about the human-machine interaction of the interface and find potential problem areas. Comparative and assessment tests instead work well when evaluating a specific feature (or potentially the entire interface), looking for specific usability problems, the main difference between the two being if a comparison is being performed. Verification tests were not studied during testing, but it is still possible to determine that these tests would only be suitable with the evaluation methods yielding quantitative and objective data.

How long a test should be, as with any usability test, will always be a trade-off between the amount of data that needs to be collected and the effort required to analyze it. The true upper bound for an eye-tracking test is determined by the equipment's limitation in storage and battery time. For Tobii Pro Glasses 3, repeated use showed that no more than one hour of recording can be expected from the battery. Other than this, an increase in test length also affected the analysis time for some evaluation methods. Evaluating longer gaze replays will inevitably take more time and if doing a GRTA, that session will be at least as long as the test session. Transcribing and analyzing participant quotes take further resources. Participant attention span and fatigue are other upper bounds to test length to consider but did not appear to be any different from regular usability testing. The lower bounds for most evaluation methods are determined on a test-by-test basis. How much data is needed to gain insights is dependent on the scenario but can be as short as a few seconds in a very limited test. If testing for cognitive workload, the need for more data increases. When comparing cognitive workload metric averages, the data collection time should probably at least be in the order of tens of seconds. If trying to find segments of high cognitive workload, the data collection time should probably be in minutes.

The findings from testing phase one suggest that while it may be possible to use eye-tracking for a low-fidelity prototype test, this would be very dependent on the objective of the research. Due to the limitation of low-fidelity prototypes, not allowing for the same breadth of interaction, exploratory testing is less suitable. The same goes for verification testing, as this requires accurate data representative of real usage which low-fidelity system prototypes may not produce. As such, hypothesis 1 appears unsupported based on the available data. For other purposes, however, such as comparing two versions of a window in early development, a low prototype fidelity will probably be "good enough". When a high degree of interaction is not required and when the data is only used for comparison, not on its own, the use of low-fidelity is especially feasible.

The level of interaction with the interface depends not only on the goal of the test but also on which evaluation methods are to be used. The type of dynamic interfaces that the tests were made on, where the user can open, move, and close different windows freely, provide a challenge for head-mounted eye-tracking. Assisted mapping, while a valuable tool, does not perform well under free interaction conditions, and even if it did, there

would be no way to adapt a still image to the changing layout of the interface. As such, when planning to use visualizations, and to some extent AOIs, a lot of thought must be put into how the test is designed to allow their effective use. Important to note, is that restricting the participants' interaction in the interface can negatively affect the validity of the data. Several participants even expressed openly that it felt unnatural.

Using experts as participants instead of non-experts in eye-tracking tests appears to yield both advantages and disadvantages. On the one hand, experts can, even in cases where a deep understanding of the interface is not required, choose to act differently than non-experts. On the other hand, experts could potentially be a liability when testing a new version of something, as their familiarity with the current version could affect the result. Using the cognitive workload metrics, it was difficult to determine whether the two types of participants differed in their results, mainly because of a lack of data. With many factors at play, hypothesis 2 could be dependent on the context and may be defined too generally. The results suggest that data from non-expert participants cannot always be used instead of data from experts, but hypothesis 2 remains inconclusive. Generally, there seems to be no reason to assume that the preferred experience level of the participant would differ in eye-tracking tests compared to traditional usability tests. Experts, being the representative user, should be used the most and are better suited for large-scale exploratory tests or tests where the accuracy of the objective data is of high importance. Oppositely, non-experts could be favorable if bias of the current system is seen as a negative.

The results in this study indicate that Nielsen's recommendation to use 5-6 participants for qualitative tests is reasonable, and is therefore the recommendation when performing gaze replay analysis and GRTA. In general, exploratory tests allow for fewer participants as the objective is to learn about potential usability problems and is not used as the basis for design decisions. With regard to the analysis of heat maps and scan paths, six participants also contributed to sufficient usability insights. While heat maps generated with fewer participants can contribute to usability insights, Nielsen's recommendation of 30 participants is probably more appropriate if conclusions are to be drawn from heat maps alone. For tests with quantitative data, this study showed that there are findings to be acquired without necessarily achieving statistical significance. While statistical significance would be necessary in a research setting, in a product development setting some data is better than no data. In corporate contexts, a low number of participants can be expected, especially if expert users are difficult to recruit. However, when using the data as the basis for conclusions, more participants are better. Moreover, one should check the result by calculating confidence intervals and statistical significance when using quantitative data as the basis for decision-making. So, the reality is that either one chooses the experiment design based on available participants to ensure high validity, or one performs tests and considers the possible invalidity of the results and draws conclusions thereafter.

7.2. Gaze Replays

Most of the benefits of eye-tracking seem to come from the ability to study how the participants' gaze moves in different situations, enabling researchers to draw more educated conclusions about user behavior. This can be done exploratively by watching the entire recording with gaze overlay, searching for interesting moments. Alternatively, predetermined moments can be studied to find out how the gaze moved in these moments. Quantifying the value of this method is difficult, but during the study the benefit of continuously reviewing gaze replays during the analysis work was clear.

7.3. Eye-Tracking Metrics & Visualizations

Eye-tracking metrics and visualizations have proved a valuable tool for adding to and expanding on insights provided by traditional usability methods and metrics. Because quantifying how many, and which insights were derived from usability metrics and eye-tracking metrics respectively is difficult, the comparison will only be discussed in general terms.

Eye-tracking technology, such as the Tobii Pro Glasses 3, provides a wide range of metrics as data outputs, and if one relevant to the research objective is found, it will likely complement the findings of traditional usability analysis. Having metrics that on their own can provide usability insights and be used to verify requirements is valuable. While only one such metric, time to first fixation, was used in this study, several other metrics have similar potential.

Measuring cognitive workload using eye-tracking under real-world conditions may not be as simple as presented in previous research. Values that correlate in theory, do not always agree in practice, which makes the data difficult to interpret. It is also important to determine before testing if cognitive workload is the most relevant measure. A participant can have a high cognitive workload without there being usability problems with the interface and vice versa. In the right context, however, measuring cognitive workload is highly relevant. Since cognitive workload metric values need to be compared to something to provide any insight, A/B testing is most likely the best use of this measure. The comparison does not however need to be to a different version of an interface, but could be to a different task, scenario, or feature in the same interface. Even though the cognitive workload values of test C (with restricted interaction) did not agree with expectations, they had more consistency than those of test D (with unrestricted interaction). As such, some discernible trends hint at the validity of hypothesis 3. Initial findings also show that it may be possible to identify intervals of high cognitive workload within a single recording. This would be a more exploratory way to use the measure but needs to be studied further to confirm its viability. Identifying these segments did prove more feasible in test D, but this is assessed as being due to the length of the test and not the level of interaction. Extracting segments does not appear accurate enough that it would be feasible to see which part of a 30-second test contains a high cognitive workload.

Knowing beforehand which metrics, visualizations, and AOIs are to be used, does not only assist in formulating the research objective but affects the test design directly. If assisted mapping is used, it may be necessary to limit the participants' interaction with

the interface, as changes in what is shown on a screen will affect the mapping and make interpretations of the visuals more difficult. While AOIs do not require assisted mapping and can be created to follow the video dynamically, this is likely too time-consuming for most practical applications.

Other than the downside of mapping, eye-tracking visualizations have shown to be effective tools for analysis. Both scan paths and heat maps are simple to create and can provide insights that are not otherwise obtainable through the raw data. Bee swarms were not extensively used in this study but were helpful in indicating whether the assisted mapping had been successful. In bee swarms, frames that failed to map correctly will usually show as long straight lines moving out towards the edges of the image.

The results suggest that eye-tracking metrics and visualizations can be used to gain insights not otherwise derived by usability metrics and methods. The findings therefore suggest potential support for hypothesis 4.

7.4. Gaze-cued Retrospective Think-Aloud

Allowing participants to comment on their actions after the test by performing a gaze-cued retrospective think-aloud provided a window to their reasoning and allowed the test leader to ask questions related to observations during the test. In general, the analysis of participants' verbalizations provided insight into their expectations and mental model of the interface with quotes highlighting specific problem areas. Transcriptions of the GRTA include numerous statements that can tell us about the method. From these, a few remarks can be made.

The first is the low resolution of the video recording produced by the glasses. Although the recording was displayed on a large projector screen, the resolution was not high enough to read text. This shortcoming prompted quotes like "It's a bit hard to see what one is doing there actually.". It also made it more difficult for the participant to stay in tune with the video, particularly in test D where the interaction was less strict and included various features. Difficulty in distinguishing between different windows that were used hindered the participant from following their actions and referring to thoughts related to them.

Another aspect to consider is the sound recording. Tobii Pro Glasses 3 records sound but the sound was turned off during the GRTA in this study as it risked interfering with the think-aloud. This decision made it more difficult to determine exactly what was going on in the recording at each given time. The structure of test C meant that while the participant received new instructions in between each task, there was no task-related interaction. However, without the sound to reveal this, these periods in between were at times mistaken for being a part of the test.

By analyzing the transcriptions from the GRTA it is clear that participants do not always remember the thoughts and reasoning behind their actions, or even the action or decision itself. This is one of the main disadvantages of RTA as explained in Chapter 3.6. Expressions such as "I think that's how I reasoned.", "I don't know what I was doing here." and "I think I saw in the menus that it wasn't, I'm unsure now, but I think

I saw that it wasn't active..." highlight this. Recalling from memory is a core aspect of RTA and the video and gaze point are supposed to facilitate the process. However, at times, the participant would look at the test leader as they were speaking, missing what was displayed on the screen and what was happening during the eye-tracking test. Consequently, one can assume, that they were basing their statements not on what they saw but solely on memory. Since, eye-tracking enables the recording of the participant's actions, describing these is not the main focus in GRTA, but rather the reasoning and thoughts behind said actions. The recommendation is therefore to be vigilant in pausing the recording and to probe the participant for reasoning behind actions.

The participant's ability to recall from memory is another aspect to consider, especially in relation to the potential contradiction of two procedural recommendations presented in Chapter 3.6. It is recommended to perform GRTA directly after an eye-tracking test has been performed to maximize the participant's ability to recall from memory. However, another recommendation is to not perform GRTA before eye-tracking tests as seeing one's gaze may influence following eye-tracking results. In this study, the decision was made to perform GRTA after test C before test D, and then again after test D. The risk of not remembering the first test and confusing it with the second was deemed larger than the influence that seeing one's gaze may have on subsequent eye-tracking tests. It was not discernable in the collective result what the effect of participants seeing their gaze was.

Nonetheless, it was evident that the gaze-cued stimuli helped the participants to recall. Occasionally, as a participant stopped talking and announced that they had covered what they were able to recall, the stimuli showed them something new and they were prompted to continue. This signals that GRTA most likely does provide more data than can be gained from the participant's memory alone. At times the participants referred to their gaze by iterating what they were observing on the recording, for example, "Here, I switch between checking the menus and [map element]." At other times, they elaborated on what they remember searching for, such as "I was trying to look and see if I was getting any feedback from [redacted] and [redacted]." and "I am searching for help here…". Furthermore, it was mentioned in Chapter 3.6 that the gaze point may cause distraction. This is hard to confirm or deny. However, it is evident that the participants referred to their gaze on multiple occasions and that the information gained from those verbalizations would be unobtainable otherwise unless the participant recalled their own gaze. The reference to their gaze was more frequent in test C than in test D. This may be due to the nature of the test, or the way questions were formulated by the test leader.

By analyzing the test leader's questions and prompts in combination with the following participant verbalizations, it is noticeable that the participants base their comments more on the stimuli if the statement from the test leader referred to it. For example, "What were you searching for here" led the participant to express what they were thinking and searching for based on what they saw on the video. Questions unrelated to what could be observed in the stimuli led them to more often base their comments on pure memory.

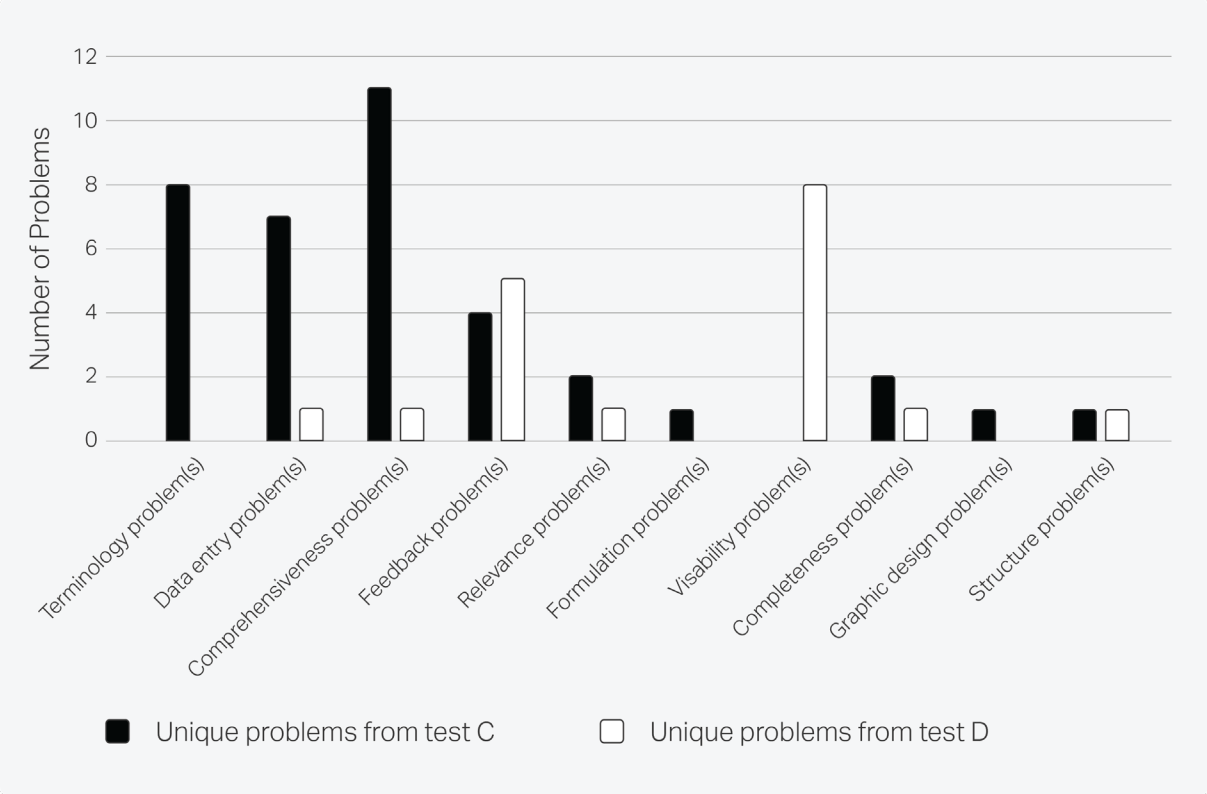
Additional remarks that can be made are that approximately half of the participants reacted to seeing their gaze with quotes like "Does one's gaze flutter like that". This corresponds to findings in the literature. Moreover, only one participant asked to pause the video. The test leader, on the other hand, paused the video on several other occasions

to allow the participant to finish their statement. Another finding is that participants requested to hear the task as it was given to them during the test again during the GRTA. This was especially evident in test C where the tasks were very specific. Only reminding them of the gist of the task was not enough for them to recall. This might be due to the level of difficulty that they experienced in this test.

The total number of unique usability problems found through GRTA differed greatly between the two tests in phase two, where 37 were expressed in test C compared to 18 in test D (see Figure 31). Test D produced half the amount of usability problems, which is more remarkable considering that test D was on average ten minutes longer. However, the higher average SUS score of 65,4 corresponds to the lower number of usability problems expressed in the GRTA (avg 4,2 per participant) in test D. The participants were tasked with an assignment that was not as hard for them to complete, hence experiencing and consequently expressing fewer problems. In contrast, the participants experienced the task in test C to be much more difficult which aligns with the low SUS score (avg 28,8) and a high average number of expressed usability problems in the GRTA (11.3 per participant).

Figure 31

Unique Usability Problems Expressed in GRTA in Test C versus Test D



Note. The number of unique usability problems expressed in the GRTA of test C versus test D for each problem type. Each specific problem is counted only once, regardless of how many participants mentioned it.

Since more exploratory tests with a wider scope, like test D, are not specifically testing a feature with known issues, it is not unreasonable to assume that it would reveal fewer usability problems. Instead, test D revealed where potential problems could exist, and laid the foundation for further studies. In test D, participants also expressed more about what they did instead of revealing as much about their thought process as in test C. This is likely due to the tasks in test D being more routine-based than in test C. Overall, even if test D did provide some insights and tests like it has a purpose, it did not provide more usability insights. The result therefore suggests that hypothesis 5 is unsupported.

08

The C2ET Method

The C2ET (Command and Control Eye-Tracking) method is developed for the use of eye-tracking in the context of usability testing on C2 interfaces at Saab Surveillance. All content in this method is either based on previous eye-tracking research or results and experience gained from user tests performed in this study. See Chapter 3 for a more detailed review of the theory and Chapter 7 for a discussion of the test results.

Although the authors have strived for high validity, the method has not been subject to evaluation in its entirety. It is moreover not clear how applicable the method is in other contexts. While some guidelines can be assumed to also apply to other complex user interfaces with expert users, this has not been tested. Although the method includes recommendations for how usability tests with eye-tracking should be designed, these are not all-encompassing, and practitioners must see to each usability test and translate the method accordingly.

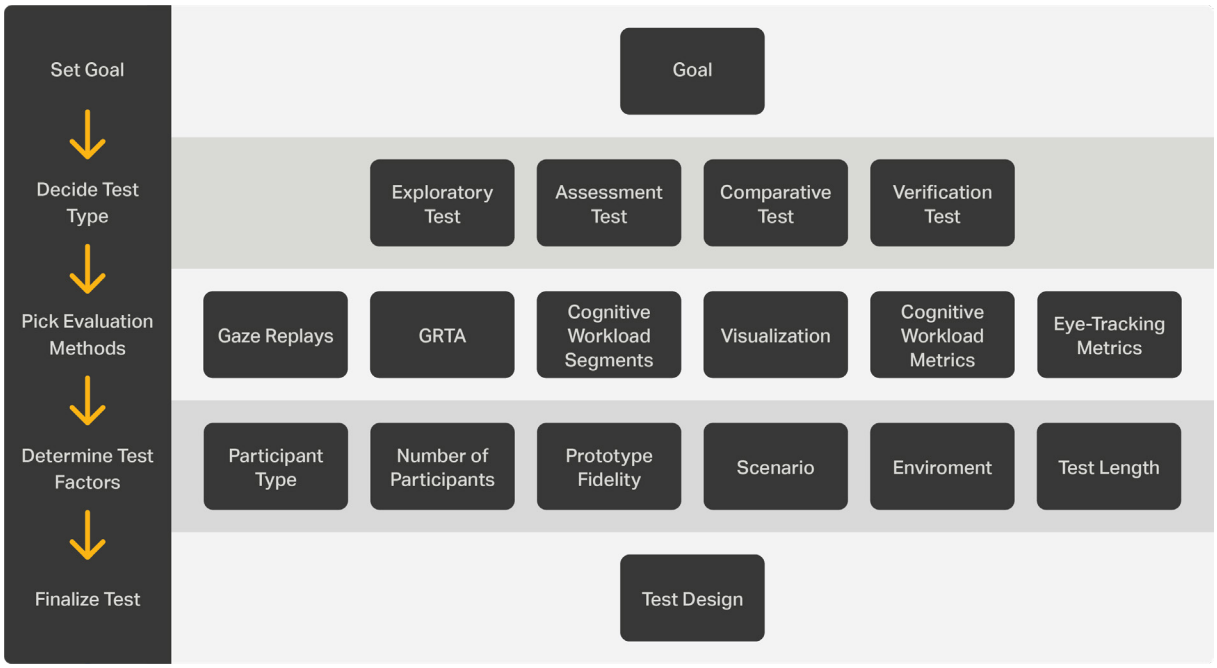
The first section of the method includes factors to consider when planning and designing eye-tracking tests for usability evaluation. Thereafter, is a description of evaluation methods and recommendations on how to analyze the eye-tracking data. The third and last section includes recommendations on preparations to be done before eye-tracking tests and a procedure with steps to complete during an eye-tracking test.

8.1. Test Design

Below is a description of different types of tests and what the recommended methods are for each of them. How long tests take, factors to consider regarding test participants, prototype fidelity, and the test environment are also presented. When designing an eye-tracking test, it is recommended to follow the order shown in Figure 32. The goal will affect what test type is recommended, which in turn will affect which evaluation methods are available. The chosen evaluation methods will affect how each test factor needs to be adapted.

Figure 32

Visualization of Test Design Workflow in C2ET



8.1.1. Type of Test

Depending on which type of test is being performed, different approaches are recommended with regard to eye-tracking. It is therefore important to choose the one most suitable for the specific context. Regardless of type, eye-tracking data should be supplemented by regular usability data, qualitative or quantitative.

The purpose of four different test types along with a recommendation for a main evaluation method and possible secondary evaluation methods as well as limitations are presented below.

Exploratory Tests

The purpose of exploratory tests is to explore the interface, learn about the human-machine interaction, and find potential problem areas.

Main Evaluation Method

Study of gaze replay – Can reveal insights about which element is processed and when.

Possible Secondary Evaluation Methods

Segments of high cognitive workload – Requires that the participant is affected by high cognitive workload during the test.

Gaze-cued retrospective think-aloud – Due to the openness of the test, is not guaranteed to contribute substantially. Requires something specific to be studied.

Visualizations – Will either be of limited value or impose restrictions on the interaction level with the interface.

Limitations

Exploratory tests are limited in their ability to gain objective data on specific problems.

Comparative Tests

The purpose of comparative tests is to compare two versions of an interface or feature, or to compare between features.

Main Evaluation Methods

Eye-tracking metrics – Which metrics are used depends on the focus of the comparison. It is not certain that relevant metrics are available.

Cognitive workload metric averages – Can indicate which of the different versions results in a higher cognitive workload for the users.

Visualizations – Can show how attention is distributed for the different versions, allowing for comparison between the two. Will require limitations in interaction for best results.

Possible Secondary Evaluation Methods

Segments of high cognitive workload – Requires that the participant is affected by high cognitive workload during the test and that the test is long enough (see section 8.2.1, Test Length).

Gaze-cued retrospective think-aloud – Can risk influencing the subsequent test if done immediately after. If there is more time between the test and the GRTA however, it can affect recall.

Study of gaze replay – Is difficult to use for comparisons.

Limitations

When doing comparative tests using eye-tracking it is important to use a within-subject test design since eye-tracking data is highly individual.

If comparing different features using eye-tracking, it is also vital to be aware of which differences in eye movement that can be attributed to the differences in the features rather than a difference in difficulty. Reading a text will for example generally have shorter fixation durations and amplitudes while marking out points on a map will generally have the opposite.

Assessment Tests

The purpose of assessment tests is to determine how efficient and effective an interface or feature performs, and to locate any potential usability problems.

Main Evaluation Methods

Study of gaze replays – Can reveal insights about which element is processed and when.

Gaze-cued retrospective think-aloud – Can provide insights into participant reasoning during the test.

Possible Secondary Evaluation Methods

Eye-tracking metrics – Which metrics are used depends on the focus of the test. It is not certain that relevant metrics are available.

Visualizations – Will either be of limited value or impose restrictions on the interaction level with the interface.

Segments of high cognitive workload – Requires that the participant is affected by high cognitive workload during the test and that the test is long enough (see section 8.2.1, Test Length).

Limitations

To get the best use of AOI metrics and visualizations, restrictions on interaction are required (due to the need for assisted mapping). To get the most realistic behavior, however, no interaction restrictions should be applied.

Verification Tests

The purpose of verification tests is to gain objective data to see if the interface passes the set requirements.

Main Evaluation Method

Eye-tracking metrics – Which metrics are used depends on the focus of the test. It is not certain that relevant metrics are available.

Limitations

Due to the need for objective data, the options for which data outputs can be used are very limited.

8.1.2. Test Length

How long a test needs to be or can be at most, depends on which evaluation methods are to be used. Task time is also crucial to the fidelity of the task scenario, with a realistic scenario potentially being up to several hours long.

Lower Bounds

The lower bounds for most evaluation methods are determined on a test-by-test basis. How much data is needed to gain any insights is dependent on the scenario but can be as short as a few seconds in a very limited test.

Depending on what is being analyzed, the lower bounds can be different:

- For cognitive workload metric averages, data collection time should at least be in the order of tens of seconds.
- For segments of high cognitive workload data collection time should be in minutes.

Upper Bounds

The true upper bound for an eye-tracking test is determined by the equipment's limitation in storage and battery time. For Tobii Pro Glasses 3, no more than one hour of recording can be expected from the battery. Other than this, an increase in test time also affects some evaluation methods:

- If studying gaze replays, a longer recording will take longer to analyze.
- If doing GRTA, the retrospective will take at least as long as the test session. Transcribing and analyzing a longer session will take further resources.

Participant attention span and fatigue are other upper bounds to test time to consider but are no different from normal usability testing.

8.1.3. Test Participants

Recommendations related to participants, i.e. level of experience, number of participants, and other criteria, are presented below.

Level of Experience

The recommendation for the level of experience of the participants generally concurs with what would be appropriate in traditional usability testing. For large-scope exploratory tests where experts can be expected to behave differently from non-experts and for verification where the accuracy of the objective data is of high importance, the use of experts is recommended. Oppositely, if new conventions are being compared to something in the current system, experts can be considered biased and should not be used. Since experts are the representative users of C2 interfaces, it would however be advised to use them for most tests.

Number of Participants

The recommendation for the number of participants differs depending on what type of data you are collecting. In addition, the type of users one wants to include and the access to such users may be a constraining factor. So, the reality is that either one chooses an experiment design based on available participants to ensure high validity, or one performs tests and considers the possible invalidity of the results and draws conclusions thereafter. Most eye-tracking tests with more than five participants produce valuable insights but the validity of the data should be questioned especially when the data is used as the basis for decision-making.

The following number of participants is recommended:

- 5-6 participants: When performing gaze replay analysis, GRTA, and exploratory heatmap, and scan path analysis.
- >5 participants: When analyzing quantitative data such as eye-tracking metrics and cognitive workload. Five participants can produce valuable insights but will probably not generate statistically significant results.
- 30 participants: When heatmaps are aggregated into a single image and are the main deliverable and basis for decision-making.

Participant Criteria

The following should be considered to ensure the quality of eye-tracking recordings:

- Inform participants to avoid wearing heavy eye makeup or big eyelashes.
- Avoid participants whose pupils are constantly dilated (e.g. because of medication).
- Exclude participants who have had eye surgery.
- Exclude participants with cataracts, amblyopia, strabismus, nystagmus, and eyelid ptosis.

8.1.4. Prototype Fidelity

Low-fidelity prototypes are not optimal if the aim is to obtain representative data, however, it can be useful under certain circumstances.

Low-fidelity prototypes can be used when:

- the aim is to study search and detection patterns
- the aim is to compare early designs
- a high degree of interaction is not required

Low-fidelity prototypes should be avoided:

- when the data is used on its own and not in comparison
- in exploratory tests
- in verification tests

The general recommendation is to perform tests on high-fidelity prototypes to ensure that the interaction in the prototype mimics the interaction in the real interface.

8.1.5. Environment

The following environmental considerations should be made to ensure the quality of the Tobii Pro Glasses 3 eye-tracking:

- Ideal illuminance is around 300 lux.
- Avoid sources that produce flickering light.
- Sunlight and devices that emit near-IR light (e.g., halogen lamps) can disrupt eye-tracking.

8.2. Evaluation Methods

Eye-tracking data should be triangulated with other data sources to cross-validate it and to understand the cognitive processes behind the data. Subjective usability metrics such as the system usability scale can be used, as well as objective usability metrics, for example, number of clicks, errors, and unnecessary actions. Moreover, gaze-cued retrospective think-aloud complements the eye-tracking recording by allowing the participant to comment on the reasoning behind actions. Below is a description of evaluation methods that can be used with eye-tracking and recommendations on how to apply them.

8.2.1. Study of Gaze Replays

Studying a gaze replay is a simple and effective way of analyzing eye-tracking data. This analysis is subjective, and it is up to the usability practitioner to determine what does and does not constitute a finding.

How to Perform a Study of Gaze Replays

Gaze replay analysis should be performed continuously when interpreting eye-tracking data to connect metrics with the user's actions in the recording. No special preparations or considerations are needed before or during testing to perform this evaluation. However, if the objective is to only analyze certain times of interest in the recording, these times of interest need to be identified through some other method.

The study of a gaze replay is done through the following steps:

1. Play the recording from the beginning.
2. Observe any usability problems or interesting behavior by looking at what the user is doing, what the user is looking at, or both combined.
3. Note down any findings with a description and time stamp.

By doing this, it is possible to observe the entire test through the eyes of the participant and find out more about what they saw and what they missed. Alternatively, if watching the entire recording is too time-consuming, these steps can be followed:

1. Find a time of interest by some other means (for example: cognitive workload segments, GRTA, or interview).
2. Play the recording at the specified time.
3. Observe any usability problems or interesting behavior by looking at what the user is doing, what the user is looking at, or both combined.
4. Note down any findings with a description and time stamp.

This is a good alternative if a problem has already been revealed and the objective is to find out more about why the problem exists.

8.2.2. Gaze-cued Retrospective Think-Aloud

Performing a GRTA is a think-aloud method performed posttest to gain insight into the participant's reasoning. This type of evaluation is subjective, both with the participant expressing their thoughts and the researcher interpreting and compiling the result.

How to Perform a GRTA

GRTA is recommended to be performed after all eye-tracking recordings have been completed so as to not risk the participant becoming too aware of their gaze. However, if more than one eye-tracking test is being performed and is to be analyzed through GRTA then the recommendation is to perform the GRTA right after each test to maximize the participant's ability to recall.

The following list includes the steps to perform a GRTA:

1. After the test has been performed, prepare to show the participant the gaze replay. If possible, use a larger screen that makes it easy to see what is happening on screen. Bear in mind that the resolution of the recording may not allow for the reading of text.
2. If it is the participants first time performing a GRTA, show a few seconds of recording and explain what they are seeing.
3. Explain to the participant that the entire test will be shown, and their task is to continuously talk about what they remember thinking in each moment. Inform them that it is not them or their performance that is being evaluated. They should be allowed to pause the recording at any time to say more. If not, important findings can be missed.
4. Use a screen recorder with audio recording to record the GRTA to be able to associate what the participant is saying with different moments. If unavailable, use a video camera recording the screen, or a normal voice recorder.
5. During the GRTA, pause the recording if the participant talks for too long about something without pausing themselves.
6. During the GRTA, prompt the participant with prepared prompts or questions if the participant stops talking. It is also possible to ask specific questions about what is seen on screen.

Examples of possible prompts:

- Keep expressing what you were thinking.
- What were your thoughts at this moment?
- Do you know what you were thinking here?
- What were you looking for here?
- Can you explain why you…
    - …clicked there?
    - …looked there?
- What did that component tell you?
- Was this what you were expecting?

To analyze the result of a GRTA, the following steps can be followed:

1. Transfer audio recordings into an intelligent verbatim transcript.
2. Extract, categorize, and compile relevant quotes as with other interview data.
3. Categorize quotes into usability problems.

8.2.3. Eye-Tracking Metrics

Eye-tracking metrics are the objective data that comes from the recording. Depending on the focus of the test, different metrics will be relevant. Deciding which metrics and possible AOIs to use in the analysis is part of the hypothesis and is a good way to understand what is to be studied. If areas of interest (AOIs) are to be used, these are also defined by the hypothesis, and redefining them post-test means changing the hypothesis. Although not technically necessary, it is good practice to define beforehand which metrics and AOIs are to be used.

General information about AOIs:

- The creation of an AOI allows the output of AOI-specific metrics that can provide further insights.
- An AOI is an area that is of special interest to the usability practitioner and needs to be studied more closely. For example:
    - an object on a map
    - a certain window in an interface
    - a picture within a window
- How small an AOI can be is limited by the accuracy of the eye-tracking device and the distance from the interface. Creating an AOI which is too small will result in unusable data.
- AOIs are defined by the hypothesis, redefining them post-test will change the hypothesis.

How to Use Eye-Tracking Metrics

How eye-tracking metrics should be used is ultimately determined by the objective of the study. The metrics from a single participant or a whole group of participants can be studied. How a metric changed over time, or averages for a whole task or test can be studied. Analysis and comparisons of the data can be done both directly in Tobii Pro Lab and exported to a spreadsheet application. Table 12 presents some common metrics and how they can be used.

Table 12

Common Eye-Tracking Metrics

Number of fixations	Can be used to measure search efficiency.
Number of fixations (AOI)	Can be used to measure how noticeable or interesting an element is.
Time to first fixation (AOI)	Can be used to measure reaction times or how noticeable an element is.
Duration of glance (AOI)	Can be used to measure how long an element is observed which can indicate how interesting it is.
Number of glances (AOI)	Can be used to measure how many times an element is observed.

Note. Examples of common metrics, as well as examples of how to use them. Metrics that can be used to indicate cognitive workload will be listed in the section about cognitive workload metrics.

To define an AOI, the following steps can be followed:

1. Define the aim of the test.
2. Choose one of the following alternatives:

    a. Use assisted mapping. This will map the gaze onto a still image allowing you to define the AOI in the image. Doing this will likely result in needing to limit the participant interaction with the interface, making sure that the layout of the stimuli does not change during the test. The still image can either be a still frame of the recording or a separate photo.

    b. Use dynamic AOIs. This will allow AOIs to be created and move around in the video as the interface changes. Tracking the AOIs onto the video will however be very time-consuming.

3. Find areas in the interface, reasonable in size, and containing visually similar elements that are relevant to the test aim. This can be done with the help of experts on the interface. AOIs should not overlap to avoid the same fixation showing up multiple times in the data.

8.2.4. Cognitive Workload Metrics

In addition to the applications previously presented, eye-tracking metrics can also be used to indicate the cognitive workload of the user. The most common metrics for this are listed in Table 13.

Table 13
Common Cognitive Workload Metrics

Fixation duration	A higher duration indicates difficulty processing information. A lower fixa-tion duration can instead indicate that there are too many stimuli to process. So, depending on the task and the definition of cognitive workload, the result may differ.	Differences in duration are usually in the order of 50-100 ms.
Pupil diameter	A higher pupil diameter indicates a higher cognitive workload. NOTE! This metric is highly affected by environ-mental conditions and should only be used if the lighting conditions have stayed consistent, nothing in the in-terface has changed, and the partici-pant only has looked at the same element.	Differences in diameter are usually in the order of 0.3-0.5 mm.
Peak saccade velocity	A lower velocity indicates a higher cognitive workload.	Differences in velocity are usu-ally in the order of 50-100 °/s.
Saccade amplitude	A lower amplitude indicates a higher cognitive workload. NOTE! This metric is correlated with peak saccade velocity.	Differences in amplitude are usually in the order of 1-5 °.

Note. Examples of common metrics for measuring cognitive workload, a description of how they correlate with cognitive workload, as well as the usual differences in order of magnitude measured between tasks. The usual differences in values are noted based on averages, individual values will be further apart.

How to Use Cognitive Workload Metrics

Since values are individual and different participants can have different baselines for the same task and environment, comparison should always be done on a within-subject basis. This is also why a value on its own does not provide any insights until it is compared to something else. One possibility is to record a baseline measurement, where the participant has no task and is only looking around. It is important to note that not all tasks will cause a high cognitive workload.

Cognitive workload metrics can be used in two ways, both require the data to be exported to a spreadsheet or coding software:

- Cognitive workload metric averages allow for comparison between versions, features, or tasks. This is done by taking the average of each metric value over the entire interval and comparing it with another interval.
- Segments of high cognitive workload allow for the identification of intervals in a recording that may have an increased cognitive workload. This is done either by creating a graph of the metrics from a single recording and visually identifying higher and lower values that coincide, or by using software to automatically extract such intervals based on threshold values. A threshold of at least the average value is needed, but a higher value is recommended.

8.2.5. Visualizations

Eye-tracking visualizations are a common way of analyzing or presenting data. Two commonly used visualizations exist, scan paths and heat maps, but Tobii Pro Lab also provides a visualization called a bee swarm. Scan paths consist of circles (representing fixations) connected by straight lines, showing the path that the participant's gaze has taken through the stimuli. Heat maps consist of a color gradient (representing few to many fixations) that shows where the participants have fixated the most. A bee swarm is a less common visualization method showing the entire path of the participant's gaze as a single continuous line.

How to Use Visualizations

To use these visualizations, the recording from the glasses must be mapped onto a still image. The still image can either be a still frame of the recording or a separate photo. While it is possible to map an entire screen, unless the screen stays the same, the visualization will not show what was shown at each respective moment. As such, it will likely be necessary to limit the interaction to make sure the contents stay similar.

Scan Paths

While it is possible to overlay the scan paths of several participants, it is usually advised to present them one at a time to avoid cluttering. The several uses of scan paths include:

- manual analysis to find or explain potential issues
- control of the quality of the data
- presentation of data
- mediating tool for doing a retrospective analysis with the participant

Things to look for in scan paths are:

- irregular scan paths can indicate that the user deviated from a regular scan cycle
- large spatial density (spread of fixation points), can indicate that the user has an indirect search approach and does not know where to look
- high frequency of transition between AOIs, can indicate that the user has an inefficient search

Heat Maps

In contrast with scan paths, it is advised to include as many participants as possible in the same heat map, to see overarching trends. To be able to aggregate the data of many participants and to compare that to how the developer thinks it should be is the main strength of the heat map. As with scan paths, it is also a good tool for presenting data.

Bee Swarm

Bee swarm can be used in video format to show multiple participants at once but can also be presented as a still image.

The several uses of bee swarms include:

- representing the gaze path with higher accuracy than what is given by a scan path, which is needed with for example reading; and
- checking the quality of the assisted mapping, as frames that failed to map correctly will usually show as long straight lines moving out towards the edges of the image.

8.3. Preparations & Procedure

The Tobii Pro Glasses 3 User Manual (Tobii, 2024a) should be advised for proper usage of the eye-tracking device.

8.3.1. Preparations

The following preparations should be made before performing an eye-tracking test.

Define Events in Tobii Pro Glasses 3 Controller Software

The Tobii Pro Glasses 3 controller application allows you to create event types ahead of the test. These can then be placed during the test. These are used in particular during analysis but can also be used to guide a GRTA. Examples of events are:

- task start, when the participant begins performing task
- task end, when the participant is finished with task
- error, when the participant performs an error

Prepare Data Collection Consent Form

Each participant in an eye-tracking test must consent to their data being collected and stored. The data collection consent form should clearly state:

- why data is being collected
- what type of data is being collected
- for how long the data is stored
- the right to withdraw consent

Calibration Validation

A calibration validation should include points for the participant to fixate on, one at a time. It should be recorded after the glasses have been calibrated. If the calibration validation shows dissatisfying quality, the glasses should be recalibrated before the test begins.

8.3.2. Materials

- Tobii Pro Glasses 3
    - Head unit
    - Recording unit
    - Ethernet cable (if not using a wireless connection)
    - SD card with adequate storage space
    - Fully charged batteries
    - Calibration cards
    - Nose pads
- Tobii Pro Glasses 3 corrective lens inserts
- Computer
- Pen
- Data consent form

If needed:

- Audio recorder
- Video recorder
- Screen recorder application
- Camera

8.3.3. Work Distribution & Roles

At least two practitioners are recommended when performing eye-tracking tests.

Test leader: instructs and guides the participant through tasks similarly, to traditional usability tests.

Eye-tracking technician:

- ensure the glasses are well mounted on the participant
- ensure all cables are in place
- place events during ongoing test
- ensure the recording is not interrupted at any time during the test

8.3.4. Procedure

The following is a procedure for eye-tracking tests. One should apply the general good practices of traditional usability tests in addition to the recommendations below.

Before Test

1. Decide on usability objective.
2. Decide on the test type.
3. Decide on evaluation method(s).
4. Decide on eye-tracking metrics and visualizations (only for specific evaluation methods).
5. Decide on AOIs (only for specific evaluation methods).
6. Decide on the experience level of participants.
7. Decide on the number of participants.
8. Recruit participants according to participant criteria. Ask new participants to find out their glasses' subscription.
9. Decide on task.
10. Decide on and prepare stimuli.
11. Prepare events in the Tobii Pro Glasses 3 Controller application.
12. Prepare and print a consent form.
13. Prepare calibration validation.
14. Ensure batteries are charged in the Tobii Pro Glasses 3 recording unit and other units (audio recorder, video camera, camera).
15. Ensure that the SD card has enough storage space.

During Test

1. Welcome the participant and showcase glasses.
2. In case the participant is wearing corrective glasses; find suitable insert lenses. Allow participant to check their vision using an eye chart.
3. If necessary, nose pads should be attached to ensure the glasses fit comfortably.
4. If the participant is new to eye-tracking, describe how eye-tracking and the glasses work. Avoid going into too much detail, as that risks making the participant aware of their gaze.
5. Inform and allow the participant to consent to data collection and storage. If a questionnaire is to be filled out, now is an appropriate time.
6. When the test is ready to begin, the first step is to ask the participant to put on the glasses, place the recording unit securely, and connect the cables. It is not recommended to use any other cables than the one provided by Tobii.
7. Calibrate the glasses. Ask the participant to hold up the calibration card at arm's length and check the Tobii Pro Glasses 3 controller application for confirmation.
8. Validate the calibration. This allows the practitioner to gain recorded data of the quality of the calibration.
9. If needed, turn on the audio recorder.
10. Begin test. Inform participant it is not their performance that is being tested.
11. The eye-tracking technician should observe the recording at all times to place events and ensure that the recording is not interrupted.
12. If a system usability scale or other post-test evaluation is to be performed, this should be filled out immediately after the eye-tracking test.
13. Gaze-cued retrospective think-aloud should be performed soon after the eye-tracking test.
14. If the data is to be analyzed using visualizations, photos can be taken of the interface using a camera. Alternatively, images can be acquired by taking a screenshot of the interface from the eye-tracking recording after the test.

09

Discussion

This chapter includes a discussion about the study as a whole. Research questions, results, process and execution, ethical considerations, and future work are the topics being discussed, providing a comprehensive understanding of the research conducted and its broader implications.

9.1. Aim & Research Questions

The aim of the thesis was to research the best practice use of eye-tracking in the evaluation of command and control interfaces and determine how to interpret its data to draw conclusions about usability and by that, answer the set research questions. The aim of the study has been achieved, and RQ1-3 addressed, by studying previous research and conducting two phases of tests, each with different focuses.

> **RQ1:** How should user tests using eye-tracking be designed to evaluate the usability of command and control interfaces effectively?

Specifically, RQ1 has been addressed through the Test Design subchapter of the C2ET method. This part includes recommendations for test types, test lengths, participants, fidelity, and environment when designing tests. Based on both previous research and the findings from both testing phases, it became evident that these factors do not have a single answer that can be generally applied. Instead, these factors have been shown to affect each other and be dependent on, for example, the chosen test type.

A limitation of the study was that not all types of user tests were performed. Exploratory, comparative, and assessment user tests were performed with eye-tracking, but verification tests were not. This is a weakness in the C2ET method, as recommendations for verification tests were included regardless. The reason for its inclusion is that these types of tests are common in the development of C2 interfaces, and since eye-tracking is a source of objective data, verification testing is a valuable area of use for the technology.

> RQ2: How should user tests on command and control interfaces using eye-tracking be performed to attain useful results?

RQ2 has been addressed through the inclusion of the Preparations and Procedure subchapter in the C2ET method. The information in this part was mainly based on initial procedures found in literature, the eye-tracking manufacture guides, and the subsequent lessons learned during the first testing phase as this procedure was applied. As mentioned, all user tests in this study were performed in a controlled setting and consequently, the C2ET method does not include guidelines for field studies. Such studies likely entail challenges that would require a separate study to unveil. Nonetheless, some of the content in the C2ET method is expected to be applicable in other settings aside from the test environment of this study.

> RQ3: Which data and methods are most useful for efficiently evaluating the usability of command and control interfaces using eye-tracking?

RQ3 has been addressed through the Evaluation Methods subchapter of the C2ET method. By studying literature, it was possible to gain an initial understanding of how to analyze eye-tracking data. These findings then had to be tried and adapted to fit the context in which the C2ET method is to be applied. The evaluation methods, more than any part of the C2ET method, is an incomplete compilation. With more testing, the presented methods can be refined even further, and unincluded methods potentially added. For example, the inclusion of the evaluation type where segments of potential high cognitive workload are identified has no real basis found in the literature. This was instead something that was found to be potentially useful when comparative testing was deemed impractical. Few such segments of value were extracted, and it was impossible to determine with certainty that these segments actually corresponded to high cognitive workload. Hence, this evaluation method needs to be studied further to determine its validity and usefulness. In general, using metrics to measure cognitive workload proved more difficult to use outside the controlled studies in previous literature. While it was not deemed completely unhelpful, the usefulness for this purpose could not be determined. On the other hand, GRTA, as documented in literature, proved to be a valuable auxiliary evaluation method to understand the cognitive processes behind the participant's gaze recorded with eye-tracking.

9.2. Results

Due to the need for consistency, the C2ET method is only based on the results from a single interface. Although similar products in Saab's portfolio were not studied, care was taken to ensure that the method could be applied more generally by not including anything in the C2ET method that required that specific interface. If applying the method to other interfaces, it should be considered that factors specific to that interface may not be included. For example, testing on interfaces that are more dispersed in nature, rather than confined to a single screen, may require the consideration of other possibilities or limitations.

The C2ET method is also based on a single type of eye tracker. As such, all conclusions are based on the results from only the head-mounted Tobii Pro Glasses 3 and its associated software. Screen-based eye-tracking would be the most relevant alternative to head-mounted in this context. In cases where only the interaction with an interface is being studied, GRTA and gaze-overlay analysis would benefit from using screen-based eye-tracking technology as that would result in better image resolution thus enabling participant and test leader to see content on the interface more clearly. Moreover, the procedure of mapping the recording onto a still image would no longer be necessary. Dynamic AOIs could then be used more extensively as they no longer need to be adjusted whenever the participant moves their head, only when they move a window in the interface. However, head-mounted eye-tracking allows for recording of a user's interaction with not only an interface but a whole human-machine system including controls and communication with the crew. The flexibility that head-mounted eye-tracking permits is also an advantage for field studies.

The C2ET method has not been subject to evaluation in its entirety due to the delimitations in the scope of the study. Consequently, the method may not be completely applicable to other interfaces and eye-tracking devices. The C2ET method should therefore not be seen as a finished document but rather as a framework for continuous improvement.

The threshold for using eye-tracking in a real product development context varies depending on which evaluation methods are used. It is possible to gain valuable insight with limited effort and resources by using gaze replays, GRTA, or visualizations. However, the use of dynamic AOIs to compensate for head movement require considerably more resources than it may be worth, especially when used in exploratory tests without a clear objective. Similarly, cognitive workload metrics also requires extensive effort, while still yielding inconclusive results.

9.3. Process & Execution

This master's thesis applied an exploratory approach, focusing on investigating potential applications of the technology rather than trying to prove anything definitively. The flexible methodology allowed for a broader understanding of the technology's capabilities and limitations, which might not have been apparent through a more constrained study. It also allowed for iterative adaptation and refinement of objectives based on findings, ensuring a thorough exploration of the technology's potential applications in C2 evaluation.

Instead of questioning or confirming previous research, this study tried to implement already-gained knowledge practically. The existing research in the areas of eye-tracking, testing, and usability was studied and used as a base from which tests could be designed. In other words, the aim was not to challenge previous research, but rather to build on it, expanding to new contexts in which the knowledge could be used. The research was scrutinized, and well-established sources within respective fields were used to the greatest extent possible. However, since there is some disagreement in eye-tracking research, such as whether long or short fixations indicate a higher cognitive workload, parts of this study may be based on misguided assumptions about which previous research to reference. Relating eye-tracking data to cognitive processes is especially difficult and

contradictory sources are common. Nonetheless, by building on previous research rather than challenging it, this study has extended eye-tracking research into a new domain, laying a foundation for future studies.

For the C2ET method to be applicable, and in turn, produce user test results that are valid and representative of reality, the fidelity of the tests that it is based on also needs to be valid. Invalid user tests risks leading to deficient design of the interface. The user tests in this study are contextually representative of where the C2ET method is expected to be applied: in a controlled setting using a system simulator. However, they are not fully representative of the real user context of the interface which entails operating in an AEW&C airplane. Additionally, the conducted testing had further limitations, only using one participant for each test instead of including a whole crew simultaneously. While this is also not fully representative of the real user context, testing with several operators simultaneously is impractical due to their availability. It is also less applicable for certain test types, such as comparative tests, where the surroundings should be more controlled. Even so, these are no new challenges specific to eye-tracking. Usability findings acquired with traditional methods are influenced by the same potential deficiency in validity. And eye-tracking, as the results in this study have shown, has the potential to contribute to a deeper understanding of the usability of these interfaces. The key is to ensure that results that are acquired from eye-tracking are interpreted holistically, together with the collective knowledge of the system, and that no single metric or visualization is used as the basis for decision-making on its own.

The limited sample size in this study inevitably impacts the reliability and generalizability of the findings. The user testing conducted involved a small number of participants, reflecting practical constraints. While statistical significance was not calculated due to the small sample size, the insights gained still offer valuable contributions to understanding the use of eye-tracking within C2 development. It is however acknowledged that larger sample sizes and more extensive testing would be necessary to confirm these findings and enhance their validity.

Collaborating with a company in the defense sector means working with confidential information, as can be expected. The thesis authors have, during the project, operated under a non-disclosure agreement prohibiting them from sharing confidential information, leading to limitations during academic supervision and in what is shared in this report.

9.4. Ethical Considerations

Saab is a company that operates in the defense sector, developing technology and products that can be used for defensive purposes, but also exploited and thus causing destruction and harm. This poses an ethical dilemma to consider. By enhancing the efficiency and usability of Saab's products the development has the potential to contribute to both sides. However, it should be noted that the area of surveillance, which the result of this study will contribute to, mainly is a tool for preventing threats. Enhancing the capabilities of surveillance technology by increasing the usability for operators allows for quicker and more accurate interpretation of data and in turn more responsible decision-making and fewer errors.

Eye-tracking allows for extensive data collection, and the question of whether the collection of data is motivated must be considered. Biometric data such as eye-tracking data is considered sensitive and must be handled thereafter, as it can indirectly contain information about personality traits, emotions, physical and mental health conditions, and more. And since eye activity is not always intentional it may be difficult for users to control what information they reveal about themselves. This introduces an ethical dilemma that must be considered by ensuring that all collection of eye-tracking data is motivated and that the data itself is anonymized and stored in a secure way.

9.5. Future Work

The subsequent step, as this study concludes, is the eventual adaptation of the C2ET method into the workflow of the HFI team at Saab Surveillance. By integrating the method into their work, the HFI team can further validate and refine the approach, ensuring it is both practical and effective in evaluating usability. This iterative process will not only help in enhancing the method itself but also in accumulating a larger dataset. It is also possible for the team to attempt applying the method in the real use-context of the C2 interface, further exploring the possibilities with eye-tracking.

Since the C2ET method should be seen as a framework to continuously improve, there is also much room for future research on the subject. The method should be evaluated in its entirety to make sure that procedures and recommendations are correct. Specifically, tests using the verification test type need to be performed to determine its viability in the context. The evaluation method for identifying segments of increased cognitive workload also needs to be researched to determine if it is possible to conclude that the intervals are identified correctly. The next step would be to apply the C2ET method on a different C2 interface, a different complex interface type, or by using a different eye-tracking device to study the generalizability of the method.

10

Conclusion

This thesis project has explored how eye-tracking technology can be used to evaluate command and control interfaces, specifically in the context of Saab Surveillance. The research questions have been addressed by studying previous research on the topics of eye-tracking, testing, and usability, and by conducting practical user tests on a C2 system simulator. Through this dual approach, it was possible to integrate theoretical insights with those from testing, providing a comprehensive examination of how to apply eye-tracking in the context of C2 development. The research questions are:

RQ1: How should user tests using eye-tracking be designed to evaluate the usability of command and control interfaces effectively?

RQ2: How should user tests on command and control interfaces using eye-tracking be performed to attain useful results?

RQ3: Which data and methods are most useful for efficiently evaluating the usability of command and control interfaces using eye-tracking?

The answers to these research questions have been compiled into the C2ET method which, in accordance with the goal of the study, will be provided to the HFI team at Saab Surveillance. The method is a stand-alone document that can serve as a guide for the HFI team or anyone else attempting to utilize eye-tracking in their C2 interface evaluation. The C2ET method includes three subchapters, focusing on different stages of eye-tracking testing:

- Test Design. This subchapter focuses on aspects to consider before the test, presenting and explaining the test factors that need to be addressed. These factors are type of test, length, participants, prototype fidelity, and environment.
- Evaluation Methods. This subchapter focuses on the evaluation of the eye-tracking data and the necessary decisions and steps to be taken before, during, and after testing for each method.
- Preparations & Procedure. This subchapter focuses on preparations directly before the test, what materials to bring to the test, the roles of the test leaders, and the overall procedure and order of the test.

The purpose of the method was never to provide conclusive answers to the implementation of eye-tracking in C2 evaluation, but rather to be a framework on which future research can be built. Moreover, the authors encourage an evaluation of the C2ET method to both reveal its flaws and determine its strengths.

In conclusion, this study demonstrated the feasibility of utilizing eye-tracking for the evaluation of command and control interfaces, highlighting its potential as a complementary tool to traditional usability methods. The gaze-cued retrospective think-aloud method was particularly valuable in providing insight into users' reasoning and mental models. On the other hand, assessing cognitive workload proved more challenging than stated in previous studies. Ultimately, this study concludes that unique usability insights indeed can be obtained with eye-tracking beyond the controlled settings of traditional lab environments to offer insights relevant to real-world product development settings. The establishment of clear procedures for the integration of eye-tracking technology within Saab Surveillance enhances the accessibility to objective and visual data which can contribute to improving the usability of C2 interfaces.

References

Alberts, D. S., & Hayes, R. E. (2006). Understanding command and control. Washington, DC: CCRP Publications.

Alhadreti, O., & Mayhew, P. (2017, May). To Intervene or Not to Intervene: An Investigation of Three Think-Aloud Protocols in Usability Testing. Journal of Usability Studies, 12(3), 111-132.

Alshammari, T., Alhadreti, O., & Mayhew, P. J. (2015). When to Ask Participants to Think Aloud: A Comparative Study of Concurrent and Retrospective Think-Aloud Method. International Journal of Human Computer Interaction (IJHCI), 6(3), 48-64. Retrieved from https://www.researchgate.net/publication/281584571_When_to_Ask_Participants_to_Think_Aloud_A_Comparative_Study_of_Concurrent_and_Retrospective_Think-Aloud_Methods?enrichId=rgreq-c27a313ade1a1ed63dc73e4352a2062a-XXX&enrichSource=Y292ZXJQYWdlOzI4MTU4NDU3MTt

American Academy of Ophthalmology. (2018, December 19). Eye Health A-Z: Rods. Retrieved February 21, 2024, from American Academy of Ophthalmology: https://www.aao.org/eye-health/anatomy/rods#:~:text=Rods%20are%20a%20type%20of,sensitive%20to%20light%20than%20cones.

Andrá, C., Lindström, P., Arzarello, F., Holmqvist, K., Robutti, O., & Sabena, C. (2013, December 11). Reading Mathematics Representations: An Eye-Tracking Study. International Journal of Science and Mathematics Education, 13(2), 237-259. https://doi.org/https://doi.org/10.1007/s10763-013-9484-y

Bergstrom, J. R., & Schall, A. (2014). Eye tracking in user experience design. Elsevier.

Borys, M., & Plechawska-Wójcik, M. (2017). Eye-tracking metrics in perception and visual attention research. EJMT, 3, 11-23.

Brooke, J. (1995). SUS - A quick and dirty usability scale.

Carter, B. T., & Luke, S. G. (2020). Best practices in eye tracking research. International Journal of Psychophysiology, 155, 49-62. https://doi.org/https://doi.org/10.1016/j.ijpsycho.2020.05.010

Casula, M., Rangarajan, N., & Shields, P. (2021, December 8). The potential for working hypotheses for deductive exploratory research. Quality & Quantity, 55, 1703-1725. https://doi.org/https://doi.org/10.1007/s11135-020-01072-9

Chen, S., Epps, J., Ruiz, N., & Chen, F. (2011). Eye activity as a measure of human mental effort in HCI. Proceedings of the 16th International Conference on Intelligent User Interfaces, (pp. 315–318). https://doi.org/10.1145/1943403.1943454

Cowen, L., Ball, L. J., & Delin, J. (2002). An Eye Movement Analysis of Web Page Usability. In X. Faulkner, J. Finlay, & F. Détienne (Ed.), People and Computers XVI - Memorable Yet Invisible. (pp. 317-335). London: Springer. https://doi.org/https://doi.org/10.1007/978-1-4471-0105-5_19

Cullipher, S., Hansen, S. J., & VandenPlas, J. R. (2018). Eye-Tracking as a Research Tool: An Introductino. ACS Symposium Series, 1-9. https://doi.org/10.1021/bk-2018-1292.ch001

Di Stasi, L. L., Antolí, A., & Cañas, J. J. (2011). Main sequence: an index for detecting mental workload variation in complex tasks. Applied ergonomics, 42(6), 807-813. https://doi.org/https://doi.org/10.1016/j.apergo.2011.01.003

Duchowski, A. T. (2017). Eye tracking methodology: Theory and practice (3 ed.). London: Springer. https://doi.org/10.1007/978-3-319-57883-5

Eger, N., Ball, L. J., Stevens, R., & Dodd, J. (2007). Cueing retrospective verbal reports in usability testing through eye-movement replay. Proceedings of the 21st British HCI Group Annual Conference on HCI 2007: HCI...but not as we know it. 1, pp. 129-137. Lancaster: BCS. https://doi.org/http://dx.doi.org/10.1145/1531294.1531312

Ehmke, C., & Wilson, S. (2007). Identifying web usability problems from eyetracking data. British HCI conference 2007, (pp. 119-128). https://doi.org/10.14236/ewic/HCI2007.12

Elbabour, F., Alhadreti, O., & Mayhew, P. (2017, May). Eye Tracking in Retrospective Think-Aloud Usability Testing: Is There Added Value? Journal of Usability Studies, 12(3), 9-110.

Epelboim, J., & Suppes, P. (2001). A model of eye movements and visual working memory during problem solving in geometry. Vision Research, 41(12), 1561-1574. https://doi.org/https://doi.org/10.1016/S0042-6989(00)00256-X

Erol Barkana, D., & Açık, A. (2014, May). Improvement of design of a surgical interface using an eye tracking device. Theoretical Biology and Medical Modelling, 11(S4). https://doi.org/https://doi.org/10.1186/1742-4682-11-S1-S4

Fernandez-Lanvin, D., Gonzales-Rodriguez, M., De-Andres, J., & Camero, R. (2023, November 21). Towards an automatic early screening system for autism spectrum disorder in toddlers based on eye-tracking. Multimedia Tools and Applications. https://doi.org/https://doi.org/10.1007/s11042-023-17694-8

Fitts, P. M., Jones, R. E., & Milton. (1950). Eye movements of aircraft pilots during instrument-landing approaches. Aeronautical Engineering Review, 9(2), 24-29.

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. International Journal of Industrial Ergonomics, 24(6), 631-645. https://doi.org/https://doi.org/10.1016/S0169-8141(98)00068-7

Goldberg, J. H., Stimson, M. J., Lewenstein, M., & Scott, N. (2002). Eye tracking in Web search tasks: Design implications. Eye Tracking Research and Applications Symposium (ETRA), (pp. 51-58). New Orleans. https://doi.org/10.1145/507072.507082

Hartmann, M., Mast, F. W., & Fischer, M. H. (2015). Spatial biases during mental arithmetic: evidence from eye movements on a blank screen. Frontiers in Psychology, 6. https://doi.org/10.3389/fpsyg.2015.00012

Havanki, K. L., & Hansen, S. J. (2018). What They See Impacts the Data You Get: Selection and Design of Visual Stimuli. ACS Symposium Series, 1292, 25-52. https://doi.org/10.1021/bk-2018-1292.ch003

Holmes, T. (2019, February 28). Eye tracking study recruitment - managing participants with vision irregularities. Retrieved from Tobii: https://www.tobii.com/blog/eye-tracking-study-recruitment-managing-participants-with-vision-irregularities

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). Eye Tracking : A Comprehensive Guide to Methods and Measures. Oxford, New York, United States: Oxford University Press.

Homeland Security. (2023, March 15). Approaches to Usability Testing. Retrieved May 14, 2024, from U.S. Department of Homeland Security: https://www.dhs.gov/cx/cx-learning/usability-testing/approaches-to-usability-testing

Hyrskykari, A., Ovaska, S., Majaranta, P., Räihä, K.-J., & Lehtinen, M. (2008). Gaze Path Stimulation in Retrospective Think-Aloud. Journal of Eye Movement Research, 2(4), 1-18. https://doi.org/https://doi.org/10.16910/jemr.2.4.5

International Organization for Standardization. (2018, March). Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts.

Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., & Bailey, B. P. (2005). Towards an index of opportunity: understanding changes in mental workload during task execution. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 311-320). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1054972.1055016

Jacob, R. J., & Karn, K. S. (2003). Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. In R. Radach, J. Hyona, & H. Deubel, The mind's eye: Cognitive and applied aspects of eye movement research. (pp. 573-605). Elsevier.

Jordan, P. W. (1993, April). Consistency and Usability. University of Glasgow Faculty of Science. Retrieved from http://theses.gla.ac.uk/5577/

Just, M. A., & Carpenter, P. A. (1980). A Theory of Reading: From Eye Fixations to Comprehension. Pyschological Review, 87(4).

Keskin, M., Ooms, K., Dogru, A. O., & De Maeyer, P. (2020). Exploring the Cognitive Load of Expert and Novice Map Users Using EEG and Eye Tracking. ISPRS International Journal of Geo-Information, 9(7). https://doi.org/https://doi.org/10.3390/ijgi9070429

Klingner, J., Kumar, R., & Hanrahan, P. (2008). Measuring the task-evoked pupillary response with a remote eye tracker. Proceedings of the 2008 symposium on Eye tracking research & applications (pp. 69-72). Association for Computing Machinery. https://doi.org/10.1145/1344471.1344489

Kovesdi, C. R., Rice, B. C., Bower, G. R., Spielman, Z. A., Hill, R. A., & LeBlanc, K. L. (2015). Measuring Human Performance in Simulated Nuclear Power Plant Control Rooms Using Eye Tracking. Idaho National Lab. https://doi.org/https://doi.org/10.2172/1261061

Kröger, J. L., Lutz, O. H.-M., & Müller, F. (2020). What Does Your Gaze Reveal About You? On the Privacy Implications of Eye Tracking. Privacy and Identity Management. Data for Better Living: AI and Privacy, 576, 226-241. https://doi.org/https://doi.org/10.1007/978-3-030-42504-3_15

Laubheimer, P. (2018, February 11). Nielsen Norman Group. Retrieved May 14, 2024, from Beyond the NPS: Measuring Perceived Usability with the SUS, NASA-TLX, and the Single Ease Question After Tasks and Usability Tests: https://www.nngroup.com/articles/measuring-perceived-usability/

Lewis, C. (1982). Using the "Thinking-aloud" Method in Cognitive Interface Design. Research Report, IBM Thomas J. Watson Research Center, Yorktown Heights, NY.

Li, N., Chen, X., Feng, Y., & Huang, J. (2022, july 15). Human-Computer Interaction Cognitive Behavior Modeling of Command and Control Systems. IEEE Internet of Things Journal, 9(14), 12723-12736. https://doi.org/10.1109/JIOT.2021.3138247

Longo, L. (2018, August 1). Experienced mental workload, perception of usability, their interaction and impact on task performance. PLoS ONE, 13(8). https://doi.org/https://doi.org/10.1371/journal.pone.0199661

McNamara, A. L., Divis, K. M., & Klein, L. M. (2019). Urgently Needed: Usability and Interaction Design in Sensor Command and Control. Sandia National Lab (SNL-NM), Albuquerque, NM (United States).

Mento, M. A. (2020, June 12). Different Kinds of Eye Tracking Devices. Retrieved February 21, 2024, from Bitbrain: https://www.bitbrain.com/blog/eye-tracking-devices

Nielsen, J. (1993). Usability Engineering. Academic Press Limited. https://doi.org/https://doi.org/10.1016/C2009-0-21512-1

Nielsen, J., & Pernice, K. (2010). Eyetracking Web Usability. Thousand Oaks, California, United States: New Riders Publishing.

O'Hara, J., Higgins, J., Fleger, S., & Pieringer, P. (2012). Human Factors Engineering Program Review Model. Office of Nuclear Regulatory Research. Retrieved from https://www.nrc.gov/docs/ML1232/ML12324A013.pdf

Olsen, A., Smolentzov, L., & Strandvall, T. (2010). Comparing different eye tracking cues when using the retrospective think aloud method in usability testing. Proceedings of the 2010 British Computer Society Conference on Human-Computer Interaction, (pp. 6-10). Dundee, United Kingdom. https://doi.org/http://dx.doi.org/10.14236/ewic/HCI2010.8

Onkhar, V., Dodou, D., & de Winter, J. C. (2023, June 15). Evaluating the Tobii Pro Glasses 2 and 3 in static and dynamic conditions. Behav Res. https://doi.org/https://doi.org/10.3758/s13428-023-02173-7

Osinga, F. P. (2006). Science, Strategy and War: The Strategic Theory of John Boyd. Routledge.

Patel, R., & Davidson, B. (2019). Forskningsmetodikens Grunder: Att planera, genomföra och rapportera en undersökning (5:5 ed.). Lund: Studentlitteratur.

Pernice, K., & Nielsen, J. (2009). How to Conduct Eyetracking Studies. Fremont: Nielsen Norman Group. Retrieved from https://www.nngroup.com/reports/how-to-conduct-eyetracking-studies/

Pignoni, G., & Komandur, S. (2022). Practical Challenges in Using Eye Trackers in the Field. Human-Automation Interaction, 12, 653-662. https://doi.org/https://doi.org/10.1007/978-3-031-10788-7

Pirus, A. E., Yulin, J., Danuaji, M. F., & Sukmaningsih, D. W. (2023). Eye Tracking in Usability Evaluation of User Experience on Enrichment Apps Web for Internship Program at XYZ University. 2023 10th International Conference on ICT for Smart Society (ICISS), (pp. 1-6). Bandung, Indonesia. https://doi.org/10.1109/ICISS59129.2023.10291901.

Pluzyczka, M. (2018). The First Hundred Years: A History of Eye Tracking as a Research Method. Applied Linguistics Papers, 4, 101-116. https://doi.org/http://dx.doi.org/10.32612/uw.25449354.2018.4.pp.101-116

Ponsoda, V., Scott, D., & Findlay, J. M. (1995). A probability vector and transition matrix analysis of eye movements during visual search. Acta psychologica, 88(2), 167-185.

Poole, A., & Ball, L. J. (2006). Eye Tracking in HCI and Usability Research. In Encyclopedia of Human Computer Interaction (pp. 211-219). Idea Group Reference.

Porta, M., Ricotti, S., & Perez, C. J. (2012). Emotional e-learning through eye tracking. Proceedings of the 2012 IEEE Global Engineering Education Conference (EDUCON) (pp. 1-6). IEEE. https://doi.org/10.1109/EDUCON.2012.6201145

Pretorius, M. C., van Biljon, J., & de Kock, E. (2010). Added Value of Eye Tracking in Usability Studies: Expert and Non-expert Participants. Human-Computer Interaction: Second IFIP TC 13 Symposium, HCIS 2010, Held as Part of WCC 2010, (pp. 110-121). https://doi.org/https://doi.org/10.1007/978-3-642-15231-3_12

Redish, J. (2007, May). Expanding Usability Testing to Evaluate Complex Systems. Journal of Usability Studies, 2(3), 102-11.

Riggs, L. A., & Ratliff, F. (1951, July 1). Visual acuity and the normal tremor of the eyes. Science, 114(2949), 17-18. https://doi.org/ doi:10.1126/science.114.2949.17

Robinson, D. A. (1968). The Oculometer Control System: A Review. Proceedings of the IEEE, Proc. IEEE, 56(6), 1032-1049. https://doi.org/10.1109/PROC.1968.6455

Russo, E. J., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. Memory & Cognition, 17(6), 759-769. https://doi.org/10.3758/BF03202637

Saab. (n.d.). 9Airborne C2. Retrieved May 15, 2024, from Saab: https://www.saab.com/products/9airborne-c2

Saab AB. (2019, 09 04). GlobalEye AEW&C in coastal environment [Photograph]. Retrieved from https://brand.saab.com/point/en/saab/media-globaleye

Sauer, J., Seibel, K., & Rüttinger, B. (2010). The influence of user expertise and prototype fidelity in usability tests. Applied Ergonomics, 41(1), 130-140. Retrieved from https://doi.org/10.1016/j.apergo.2009.06.003

Sauro, J. (2018, September 19). *5* Ways to Interpret a SUS Score. Retrieved May 2, 2024, from Measuring U: https://measuringu.com/interpret-sus-score/

Schindler, M., & Lilienthal, A. J. (2019, February 6). Domain-specific interpretation of eye-tracking data: towards a refined use of the eye-mind hypothesis for the field of geometry. Educational Studies in Mathematics, 101, 123-139. https://doi.org/https://doi.org/10.1007/s10649-019-9878-z

Skaramagkas, V., Giannakakis, G., Ktistakis, E., Manousos, D., Karatzanis, I., Tachos, N. S., . . . Tsiknakis, M. (2023). Review of Eye Tracking Metrics Involved in Emotional and Cognitive Processes. IEEE Reviews in Biomedical Engineering, 16, 360-277. https://doi.org/10.1109/RBME.2021.3066072

Snyder, C. W., Mastrandrea, M. D., & Schneider, S. H. (2011). The complex dynamics of the climate system: constraints on our knowledge, policy implications and the necessity of systems thinking. (C. Hooker, Ed.) Handbook of the Philosophy of Science, 10: Philosphy of Complex Systems, 467-505.

Stieff, M., Hegarty, M., & Deslongchamps, G. (2011). Identifying Representational Competence with Multi-Representational Displays. Cognition and Instruction, 29(1), 123-145. https://doi.org/https://doi.org/10.1080/07370008.2010.507318

Svensson, J., Wikström, M., & Lif, P. (2020). Metoder för user experience - UX: Tillämpning vid utveckling av HMI för Gripen E. FOI-R--5023--SE, FOI. Retrieved from https://www.foi.se/rest-api/report/FOI-R--5023--SE#:~:text=Omr%C3%A5det%20som%20UX%20omfattar%20v%C3%A4xer,%2C%20inneh%C3%A5llsstrategi%2C%20tillg%C3%A4nglighetsanpassning%20och%20analysverktyg.

Swedish Authority for Privacy Protection. (2021, April 16). Sensitive Personal Data. Retrieved April 29, 2024, from https://www.imy.se/en/individuals/data-protection/introduktion-till-gdpr/what-is-actually-meant-by-personal-data/what-is-meant-by-sensitive-personal-data/

Tobii. (2022, September 14). How to perform manual and assisted mapping. Retrieved from Tobii Connect: https://connect.tobii.com/s/article/how-to-perform-manual-and-assisted-mapping?language=en_US

Tobii. (2023a, June 19). Understanding Tobii Pro Lab's eye tracking metrics. Retrieved from Tobii Connect: https://connect.tobii.com/s/article/understanding-tobii-pro-lab-eye-tracking-metrics?language=en_US

Tobii. (2023b, July 6). Creating good conditions for eye tracking. Retrieved 05 13, 2024, from Tobii Connect: https://connect.tobii.com/s/article/Creating-good-conditions-for-eye-tracking?language=en_US

Tobii. (2023c, March 29). Tobii eye tracker glossary. Retrieved February 21, 2024, from Tobii Connect: https://connect.tobii.com/s/article/tobii-eye-tracker-glossary?language=en_US

Tobii. (2024a, January). Tobii. Retrieved from Tobii Pro Glasses 3: User Manual: https://go.tobii.com/tobii-pro-glasses-3-user-manual

Tobii. (2024b). Tobii Pro Lab (Version 1.232.52758) [Computer software]. Retrieved from https://www.tobii.com/

Tobii. (n.d. -a). Products. Retrieved February 21, 2024, from Tobii: https://www.tobii.com/products#eye-trackers

Tobii. (n.d. -b). Tobii Pro Glasses 3 Field Guide. Retrieved April 22, 2024, from Tobii Connect: https://connect.tobii.com/s/field-guide-glasses3?language=en_US

Tobii. (n.d. -c). Calibration. Retrieved from Tobii pro/sdk: https://developer.tobiipro.com/commonconcepts/calibration.html#:~:text=The%20common%20procedure%20of%20doing,the%20participant%20was%20expected%20to

van den Haak, M. J. (2008). A Penny for Your Thoughts - Investigating the Validity and Reliability of Think-Aloud Protocols for Usability Testing.

van den Haak, M. J., De Jong, M., & Schellens, P. J. (2003, September). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. Behavior and Information Technology, 22(5), 339-351. https://doi.org/10.1080/0044929031000

Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T.-P. (2001). Eye activity correlates of workload during a visuospatial memory task. Human factors, 43(1), 111-121.

Vermeulen, J., Luyten, K., van den Hoven, E., & Coninx, K. (2013). Crossing the Bridge over Norman's Gulf of Execution: Revealing Feedforward's True Identity. CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, (pp. 1931-1940). https://doi.org/https://doi.org/10.1145/2470654.2466255

Wang, J., Antonenko, P., Celepkolu, M., Jimenez, Y., Fieldman, E., & Fieldman, A. (2019). Exploring Relationships Between Eye Tracking and Traditional Usability Testing Data. International Journal of Human–Computer Interaction, 35(6), 483-494. https://doi.org/10.1080/10447318.2018.1464776

Wang, Q., Yang, S., Liu, M., Cao, Z., & Ma, Q. (2014). An eye-tracking study of website complexity from cognitive load perspective. Decision Support Systems, 62, 1-10. https://doi.org/https://doi.org/10.1016/j.dss.2014.02.007

Xie, W., Lee, M. H., Chen, M., & Han, Z. (2023, September 01). Understanding Consumers' Visual Attention in Mobile Advertisements: An Ambulatory Eye-Tracking Study with Machine Learning Techniques. Journal of Advertising. https://doi.org/https://doi.org/10.1080/00913367.2023.2258388

Yang, C., Liu, Z., Zhou, Q., Xie, F., & Zhou, S. (2014). Analysis on Eye Movement Indexes Based on Simulated Flight Task. Engineering Psychology and Cognitive Ergonomics: 11th International Conference, EPCE 2014, Held as Part of HCI International 2014 (pp. 419-427). Heraklion, Crete, Greece: Springer International Publishing.

Yeaton, W., Langenbrunner, J. C., Smyth, J. M., & Wortman, P. M. (1995, October). Exploratory Research Synthesis Methodological Considerations for Adressing Limitations in Data Quality. Evaluation & the Health Professions, 18(3), 283-303. https://doi.org/http://dx.doi.org/10.1177/016327879501800304

Zagermann, J., Pfeil, U., & Reiterer, H. (2016). Measuring Cognitive Load using Eye Tracking Technology in Visual Computing. Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization (pp. 78-85). Baltimore, MD, USA: Association for Computing Machinery. https://doi.org/10.1145/2993901.2993908

Zhou, C., Yuan, F., Huang, T., Zhang, Y., & Kaner, J. (2022). The Impact of Interface Design Element Features on Task Performance in Older Adults: Evidence from Eye-Tracking and EEG Signals. International Journal of Environmental Research and Public Health, 19(15). https://doi.org/https://doi.org/10.3390/ijerph19159251

Appendices

Appendix A – Survey for Phase One

Please note that your participation is voluntary, and you can choose to cancel your participation at any time. All data fields are optional to fill in.

1. **Gender (circle one):**

   Male Female Other/prefer not to say

2. **Year of birth: _____**
3. **Do you have system (C2) related military service (circle one):**

   Yes No Prefer not to say

4. **Rate your level of operator experience (circle one):**

   No operator experience

   Low level of operator experience

   Medium level of operator experience

   High level of operator experience

5. **Rate your level of experience with the C2 interface (circle one):**

   No experience

   Low level of experience

   Medium level of experience

   High level of experience

6. **Do you have a current medical eye condition? (Does not include impaired vision) (circle one):**

   Yes No Prefer not to say

7. **Have you undergone eye surgery? (circle one):**

   Yes No Prefer not to say

Appendix B – Survey for Phase Two

Please note that your participation is voluntary, and you can choose to cancel your participation at any time. All data fields are optional to fill in.

1. **Gender (circle one):**

   Male Female Other/prefer not to say

2. **Year of birth: _____**
3. **Do you have system (C2) related military service (circle one):**

   Yes No Prefer not to say

4. **Rate your level of operator experience (circle one):**

   No operator experience

   Low level of operator experience

   Medium level of operator experience

   High level of operator experience

5. **Rate your level of experience with the C2 interface (circle one):**

   No experience

   Low level of experience

   Medium level of experience

   High level of experience

6. **Rate your level of experience with the [tested feature] (circle one):**

   I have never heard of the [tested feature]

   I have heard of the [tested feature] but never used it

   I have used the [tested feature]

7. **Have you been active in the development of the [tested feature]? (circle one):**

   Yes No Prefer not to say

8. **Do you have a current medical eye condition? (Does not include impaired vision) (circle one):**

   Yes No Prefer not to say

9. **Have you undergone eye surgery? (circle one):**

   Yes No Prefer not to say

Appendix C – Test Description for Phase Two

Test A

The purpose of this test is to see how well the eye-tracking metrics and visualizations can identify usability issues. This will be done by testing part of the interface statically, which means that windows cannot be moved and there is only one way to solve the task. The chosen function has been selected in part due to indications of existing usability problems and in part due to its ability to be tested statically.

The reasoning behind doing a static test is that previous tests have shown that metrics that can quantify things such as cognitive workload can be difficult to analyze in tests of greater scope. To not be dependent on the availability of different versions or prototypes of the same interface or window, the goal is to identify usability issues or validate the current version from a single version. To get a baseline measurement of low cognitive workload, this will be recorded after each test.

Three different types of data are being collected for analysis: eye-tracking metrics, traditional usability data, and input from the gaze-cued retrospective think-aloud.

Test Procedure

1. Preparations before participant arrival
2. Survey and background information
3. Corrective lenses
4. Test description
5. Calibration
6. Recording and verification
7. Start simulation
8. Test start and task descriptions
9. Execution
10. Record baseline
11. Stop recording
12. System Usability Scale
13. Retrospective Think-Aloud

Detailed Description

1. **Preparations before participant arrival**

   - Place track points around the screen (4–8 points) with a high contrast to facilitate assisted mapping in analysis.
   - Prepare a place for the eye-tracking technician to be seated from where they can observe the output from the glasses.

- Prepare a place for the test leader to be seated to give instructions to the participant.
- Start the eye-tracking computer and associated eye-tracking software. Make sure the computer is connected with a power cord.
- Verify that the eye-tracking glasses battery unit has a charged battery and sufficient memory for the test.
- Connect the eye-tracking glasses to the battery unit, which in turn is connected to the computer.
- Place an audio recorder at the test leader's position.
- Prepare the corrective lenses if needed.
- Start the system simulator.
- Open all interface windows that are needed and place them correctly.
- Prepare the survey and system usability scale papers. Verify access to pen.

2. Survey and background information

The test leader gives the participant the following information:

> We are doing a master's thesis on eye-tracking and how it can be used to evaluate or validate interfaces in the C2 context. Eye-tracking is a technology that can collect data on where and how you are looking. The eye-tracking glasses have a camera that records what you are seeing, as well as sensors that determine what you are focusing on at any given moment.
>
> Today, you will be participating in two tests and a number of evaluations in relation to that. The tests do not intend to test you or your knowledge of the interface. Regardless of results, the collected data will be useful. We will soon describe the scenario and task.

The participant is asked to fill out a survey with background information as well as consent to data collection.

3. Corrective lenses

If needed, corrective lenses for impaired vision can now be tried out for the eye-tracking glasses.

4. Test description

Here was a description of how the test was to be conducted. This section has been redacted due to confidentiality.

5. Calibration

Input participant-id and test-id into the name input in the eye-tracking software. For example, “Participant 1-A”. Observe that test-id only is necessary in the event that several recordings are made with the same participant since the software does not allow identical names.

Let the participant hold a calibration card in front of them until a green circle appears in the video stream, then press calibrate. Observe that if calibration cards are used as track points the user should not have these in the field of view when performing the calibration. If calibration fails, or succeeds but looks poor, adjust the calibration card position and try again.

6. Recording and verification

Start the eye-tracking recording.

Start the audio recording.

Say the participant-id and test-id into the recording device (to help identify the recording).

Clap (to sync different audio recordings).

When calibration has been performed, this can be verified using a prepared sequence of points that the participant is told to look at on a screen. Make sure the screen with the verification sequence is right in front of the participant. If the participants gaze is offset, the recording can be aborted and the calibration redone. Including the verification in the eye-tracking recording enables the possibility of looking back at the quality of the recording in the analysis stage.

Observe that if the recording freezes or turns black during the test, the recording will not be saved in a correct way and the test needs to be restarted or aborted.

7. Start simulation

The simulation can now be started.

8. Test start and task descriptions

[Part of this section has been redacted due to confidentiality.]

Between each task the test leader asks:

- How confident are you that you have achieved the goal of the task?
- Can you rate your confidence on a scale from 1-5, 1 being not at all confident and 5 being completely confident?

9. Execution

[This section has been redacted due to confidentiality.]

10. Record baseline

Give the participant the following information and let the recording continue for a minute.

> The test is now over. We wish that you relax and take a minute to look around in the interface without interacting. Look at different elements and windows without trying to remember anything.

11. Stop recording

Stop the recording and make sure the file is saved in a correct way.

12. System Usability Scale

Give the participant a system usability scale form, a pen and the following information.

> We now want you to fill out the following form. Mark one box on each row with your first reaction to the function you just used. Do not think too long on each question. If you are unsure or cannot answer you can mark the center box. Remember to only fill out the form based on the functions you used in this test.

Label the form with test-id, participant-id and date.

13. Retrospective Think-Aloud

Move the participant to a location suitable for a retrospective. If possible present the video on a larger screen. Open the just recorded gaze video in the Glasses 3 software.

Start screen recording with sound.

The test leader provides the following information:

> You will now see a recording of the task you just completed where you also will see what you are looking at. We will start by showing you what this will look like.

Show the beginning of the recording, stopping before task execution.

> We will now show you the entire recording and we want you to think aloud, talking us through what you did and thought while looking at the screen. We would like to remind you that it is not your performance, but the interface that we are testing. As such, there is no right or wrong answer. You can pause the recording at any time by pressing here.

Start the recording.

When a new task is given in the video, repeat the task description for the participant.

During the retrospective, prompts or specific question can be used to keep the participant talking.

Prompts:

- Keep expressing what you were thinking.
- What did you think at this moment?
- Can you tell us what you were thinking here?
- What are you looking for here?
- Do you remember what you were thinking at this moment?
- Can you tell us why you…?
    - pressed that.
    - looked there.
- What does this element tell you?
- Was this what you were expecting?

Analysis

Eye-tracking data

Use the data to find specific moments where the user experienced difficulties. Can also be used to see if known errors or insecurities can be confirmed by the data.

- Metrics – Look for high cognitive workload or inefficient search in specific moments.
    - Fixation duration (binned)
    - Fixation frequency (binned)
    - Saccade amplitude (binned)
    - Peak saccade velocity (binned)
    - Pupil diameter (binned)
- Visualizations
    - Scan path – Look for inefficient scan path or bad search efficiency.
    - Heat map –See which areas are looked at the most. Together with other data, try to determine why. Could it be because of confusion or interesting elements?
    - Bee swarm – Check quality of the assisted mapping.

Usability data

- Objective data – Quantify the usability of the interface
  - Number of clicks (compare to optimal number)
  - Number of errors (also compare to number of actions)
  - Number of unnecessary actions
  - Task completion time
  - Has the task goal been achieved? Yes/No
- Subjective data
  - Interview questions – Increase knowledge about user behavior
  - System Usability Scale – Quantify the usability of the interface

Gaze-cued retrospective think-aloud

- User comments – Increase understanding of user actions

Test B

The purpose of this test is to see how well retrospective think-aloud works in the context of evaluating the C2 interface. This will be done by testing part of the interface dynamically, which means that there are less restrictions on how the participant can interact with the interface compared to test A. As such, there can be several ways for the participant to achieve the goal of the task. Several functions have been chosen as the focus of the test due to their similarity of purpose and difference in execution.

The reason behind doing a test dynamically is to put the task in its context of use and let the participants solve the task on their own in their own order, similarly to how it would have been done in a real case. Like in test A, there is no comparison between different versions of the interface or interface elements. Some comparison will however take place between the different functions.

Because of the more open nature of test B, it is uncertain whether the objective eye-tracking metrics will yield any insights. These metrics will however be analyzed with the purpose of determining their value. To get a baseline measurement of low cognitive workload, this will be recorded after each test.

Three different types of data are being collected for analysis: eye-tracking metrics, traditional usability data, and input from the gaze-cued retrospective think-aloud.

Test Procedure

1. Preparations
2. Task description
3. Calibration
4. Recording and verification
5. Start simulation and test
6. Execution
7. Record baseline
8. Stop recording
9. System Usability Scale
10. Retrospective Think-Aloud
11. End of test

Detailed Description

1. Preparations

- Prepare the interface for the test by removing everything added in test A.

2. Task Description

Part of this section has been redacted due to confidentiality.

The test leader now provides the participant with a memory aid for the task.

3. Calibration

Input participant-id and test-id into the name input in the eye-tracking software. For example, “Participant 1-B”. Observe that test-id only is necessary in the event that several recordings are made with the same participant since the software does not allow identical names.

Let the participant hold a calibration card in front of them until a green circle appears in the video stream, then press calibrate. Observe that if calibration cards are used as track points the user should not have these in the field of view when performing the calibration. If calibration fails or succeeds but looks poor, adjust calibration card position and try again.

4. Recording and verification

Start the eye-tracking recording.

Start the audio recording.

Say the participant-id and test-id into the recording device (to help identify the recording).

Clap (to sync different audio recordings).

When calibration has been performed, this can be verified using a prepared sequence of points that the participant is told to look at on a screen. Make sure the screen with the verification sequence is right in front of the participant. If the participants gaze is offset, the recording can be aborted, and the calibration redone. Including the verification in the eye-tracking recording enables the possibility of looking back at the quality of the recording in the analysis stage.

Observe that if the recording freezes or turns black during the test, the recording will not be saved in a correct way and the test needs to be restarted or aborted.

5. Start simulation and test

Start simulation.

The test leader informs the participant:

> The test has now begun, and you can begin solving the task.

6. Execution

Part of this section has been redacted due to confidentiality.

After the task the test leader asks:

> How confident are you that you have achieved the goal of the task?
>
> Can you rate your confidence on a scale from 1-5, 1 being not at all confident and 5 being completely confident?

7. Record Baseline

Give the participant the following information and let the recording continue for a minute.

> The test is now over. We wish that you relax and take a minute to look around in the interface without interacting. Look at different elements and windows without trying to remember anything.

8. Stop Recording

Stop the recording and make sure the file is saved in a correct way.

9. System Usability Scale

Give the participant a system usability scale form, a pen and the following information.

> We now want you to fill out the following form. Mark one box on each row with your first reaction to the function you just used. Do not think too long on each question. If you are unsure or cannot answer you can mark the center box. Remember to only fill out the form based on the functions you used in this test.

Label the form with test-id, participant-id and date.

10. Retrospective Think-Aloud

Move the participant to a location suitable for a retrospective. If possible present the video on a larger screen. Open the just recorded gaze video in the Glasses 3 software.

Start screen recording with sound.

The test leader provides the following information:

> We will now show you the entire recording and we want you to think aloud, talking us through what you did and thought while looking at the screen. We would like to remind you that it is not your performance, but the interface that we are testing. As such, there is no right or wrong answer. You can pause the recording at any time by pressing here.

Start the recording.

During the retrospective, prompts or specific question can be used to keep the participant talking.

Prompts:

- Keep expressing what you were thinking.
- What did you think at this moment?
- Can you tell us what you were thinking here?
- What are you looking for here?
- Do you remember what you were thinking at this moment?
- Can you tell us why you…?
    - pressed that.
    - looked there.
- What does this element tell you?
- Was this what you were expecting?

11. End of Test

Thank the participant for their participation.

Analysis

Eye-tracking data

- Use the data to find specific moments were the user experienced difficulties. Can also be used to see if known errors or insecurities can be confirmed by the data.
- Metrics – Look for high cognitive workload or inefficient search in specific moments.
    - Fixation duration (binned)
    - Fixation frequency (binned)
    - Saccade amplitude (binned)
    - Peak saccade velocity (binned)
    - Pupil diameter (binned)
- AOI metrics (use interface windows as AOIs) – See if the cognitive workload differs between windows.
    - Total duration of fixations
    - Average duration of fixations
    - Average pupil diameter
    - Number of fixations
    - Fixation frequency
- Visualizations
- Scan path – Look for inefficient scan path or bad search efficiency.
- Heat map –See which areas in each window are looked at the most. Together with other data, try to determine why. Could it be because of confusion or interesting elements?
- Bee swarm – Check quality of the assisted mapping.

Usability data

- Objective data – Quantify the usability of the interface
    - Number of clicks (compare to optimal number)
    - Number of errors (also compare to number of actions)
    - Task completion time
    - Has the task goal been achieved? Yes/No
- Subjective data
    - System Usability Scale – Quantify the usability of the interface

Gaze-cued retrospective think-aloud

- User comments – Increase understanding of user actions

DEPARTMENT OF INDUSTRIAL AND MATERIALS SCIENCE
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2024
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY