Investigating the Influence of Display Size on Cognitive Load in Visual Search Tasks using Eye-Tracking Technology

Master thesis

by

Javid Guliyev

at the

Universität Konstanz

Department of Computer and Information Science

1st Referee: Prof. Dr. Harald Reiterer

2nd Referee: Prof. Dr. Marc H. Scholl

Konstanz, 2018

Abstract

Over the past few years, researchers are showing interest to study the level of cognition with the modern eye tracking technologies which can accurately measure gaze movements. Rise of the eye tracking technology encouraged researchers to use cognitive load as another measurement in the field of Human-Computer Interaction as well.

In this thesis, we use an eye tracking technology as a tool to assess the influence of the display size on the cognitive load in visual search tasks. We designed and conducted an empirical study with 35 participants to investigate if the display size has any effect on the level of cognition. First, we analysed how gaze measurements were affected while changing the level of cognition in two different sized displays. Later, we investigated if the display size had any influence on the gaze measurements. Using this knowledge, we discussed how the display size influenced the level of cognition in visual search tasks.

During the experiment, we collected not only the gaze measurement, but also the level of performance and users' subjective thoughts about their mental effort.

Our objective and subjective analysis shows that display size has an influence on the cognitive state and there is a tendency for the large display leading to a higher cognitive load in visual search tasks.

Table of Contents

List of Tables

List of Figures

1. Introduction

In addition to the usability measurements such as effectiveness, efficiency and satisfaction, the cognitive load became another measurement to assess human task performance [1].

Electroencephalography (EEG) and magnetoencephalography (MEG) are traditional methods to measure cognitive state of the human. However, one needs to have guidance from neuroscientist to use these methods. The eye-tracking technology, on the other hand, allows the researchers not only from neuroscience but from different areas to assess the cognitive state of the human.

Choosing the right display size to create efficient visual interface to support users' cognitive ability is an important issue in HCI. Thanks to the modern technology, we can also assess the level of cognition in frame of HCI and design cognitively less demanding interfaces.

Based on aforementioned, we come up with an idea to analyse the influence of the display size on the cognitive load and defined following research question:

How does the Display Size Influence Cognitive Load in Visual Search Tasks?

Many researchers investigated the influence of the display size on the users' performance and their cognitive state for different tasks. However, there is no research conducted to investigate the influence of the display size on the cognitive load using eye-tracking technology.

The lack of research on this direction motivated us to design a study, conduct an experiment and analyse the influence of the display size on the cognitive load in visual search tasks using eye-tracking technology.

1.1 Thesis outline

In Chapter 2, we present the theoretical background related to our thesis. First, we will define and discuss the cognitive load and its measurement methods. In the next section, we will introduce eye tracking technology and its different measurements.

In chapter 3, we present related works conducted to assess the cognitive load with eye tracking technology and the influence of the display size in HCI.

In chapter 4, we define the research question and research sub-questions which will help us to answer our main research question.

In chapter 5, we describe the pilot study conducted before the large-scale experiment. We mentioned the shortcomings of our initial study design and this will help us to improve our experimental design.

Chapter 6 covers the large-scale experimental design. We will present the task description and implementation, apparatus which are used during the experiment, participants and the whole procedure to conduct the experiment.

Chapter 7 provides detailed information about the data analysis. We will describe the collected data and applicability of the different methods to analyze our data.

In chapter 8, we present the results of the analysis for each independent variable separately.

Chapter 9 covers the discussion of the results presented in Chapter 8. We will discuss the applicability of different gaze event to assess the level of cognition in different displays and the influence of the display size on the cognitive load.

In Chapter 10, we present the conclusion of the research and new ideas for the future investigations.

2. Theoretical background

In this thesis, we measure the influence of the display size on the cognitive load during the visual search and make use of eye-tracking technology to conduct our research. The cognitive load and eye-tracking are the main terms which will be used throughout the thesis. Therefore, in this chapter, we will briefly introduce these terms.

In the first section, we will define the cognitive load, its types and the methods to assess it. In the second section, we will discuss the different eye-tracking technologies, gaze movements and the gaze events which are relevant to our research.

2.1 Cognitive load

Cognitive load is the amount of the mental effort used in the working memory [2]. Working memory is the cognitive system and its main mission is to hold information for processing [3]. Baddeley et. al [4] proposed the model of the working memory. Their theory claims that the working memory has three components: the central executive, the phonological loop, and the visuospatial sketchpad.

The phonological loop stores the phonological information such as the sound of the spoken language. If one calls the phone number with a high number of digits out several times, then the phonological loop stores this information [5]. On the other hand, visuospatial sketchpad stores the visual information. The central execute coordinates these two components and it is responsible for directing the attention of each component to the relevant information.

Cognitive load is sometimes called a mental effort or a task load to describe the mental state during problem-solving [6].

The capacity of the working memory is limited, accordingly the ability to process the information is restricted [7]. Therefore, in HCI it is important to provide the information in such a way that users' task performance is improved.

Measuring cognitive load is an important issue of the cognitive theory. Electroencephalography (EEG) and magnetoencephalography (MEG) are traditional methods to measure the cognitive load. [6]. These methods are used to measure the brain activity. Very sensitive electrodes or magnetometers are placed on the human scalp and these receivers can detect electrical or magnetic fields produced by the neurons of the brain [8][9]. However, these methods are expensive and need a professional assistance.

Furthermore, the cognitive load can be measured by methods such as blood pressure, heart rate, facial expressions and gaze movements [10]. In our research, we use the gaze movements and pupillary responses to measure the cognitive load.

2.2 Eye tracking

Eye tracking describes the process of capturing the eye activity. An eye tracker is a device which is measuring the position of eye movements and records different gaze events. Most modern eye trackers use infrared technology to reflect the light on the cornea which is called *pupil center corneal reflection* (PCCR) [11](see Figure 1). Additionally, this kind of eye trackers uses high-resolution cameras to track the eye movements [12].

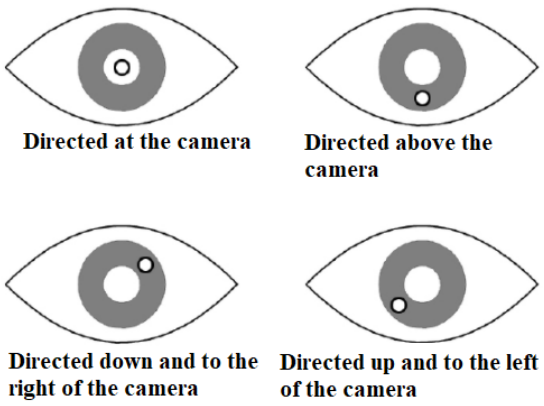


Figure 1. The reflection of the infrared light on the cornea [13]**.**

There are two types of eye trackers: *screen based-eye trackers* (sometimes are called remote eye tracker) and *mobile eye trackers*. Screen-based eye trackers are usually suited in front

of the monitor and record the eye movements within certain limits [12] and during the eye tracking process, the user cannot freely move around. Mobile eye trackers, on the other hand, gives an advantage of moving around. These kind of eye trackers are useful if the experiment has to be conducted in the natural environment.

For our experiment, we used an eye-tracking glasses which belong to the mobile eye tracker category.

Gaze events can be categorized into two groups: voluntary and involuntary eye movements. In the following, we will describe the eye movements which are related to our research [14].

Fixation is a voluntary eye movement which is the collection of 20 to 50 gaze points based on specified area and timespan. These gaze points are recorded between 200 and 300-millisecond timespan. During the fixation event eyes remain still. Common metrics for fixations are the number of fixations, the duration of the fixation and the position of fixation as x-and y- coordinates in pixel. As we will discuss in the next chapter, there is a relation between the gaze event fixation and cognitive load.

Saccade is a voluntary eye movement from one fixation to another. It lasts 30 to 80 milliseconds which is the fastest movement the human body can perform. Common metrics for saccade are the distance which it travels, duration and the number of saccades per second.

Blink is the rapid closing of the eyelid and it is an involuntary eye movement. Human has partial control over this gaze event and sometimes blinking are described as a voluntary eye movement. Blinking is important to keep the eyes moisten and protect them.

Pupil dilation is a type of the pupillary response and it is involuntary eye movement. Pupil dilation can have several causes such as response to the light, sexual stimulation, or interest to the subject [15]. The task-invoked pupillary response is the type of the pupillary response caused by a cognitive load [16]. Pupil dilation is the most widely studied eye movement which has a direct relation to the cognitive load.

3. Related work

In this chapter, we will review related works which are relevant to our research. We will discuss previous works which are related to the measuring cognitive load with eye-tracking technology and the display size in HCI and its effects.

3.1 Eye tracking and cognitive load

The task-invoked pupillary response was and still is the most reliable and most studied eye movement which has a direct relation to the cognitive load [17]. However, modern eye tracking technologies enable to study the relation between the cognitive load and different gaze events such as fixation, saccade, and blinks. During the last years, there have been many researches which investigate the relation between eye tracking and cognitive state. In this section, we will review the most relevant researchers done on this topic.

Peysakhovich et. al [18] investigated the relation between the pupil diameter and the cognitive load. During the experiment, authors used simple piloting task with an auditory-visual interference paradigm. The participants were controlling an imaginary aircraft with the joystick. The working memory was manipulated by increasing the complexity of the task. Authors concluded that the task difficulty has significant effect on pupil diameter: the higher cognitive load leads to a larger diameter. Their results show that there is a significant relation between the cognitive state and pupil dilation.

Nourbakhsh et. al [19] used the blink rate and galvanic skin response (GSR) to measure cognitive load in real time. During the experiment, they used arithmetic task with 4 difficulty level. Authors showed that using GSR and eye tracking technology together the accuracy of the assessment of cognitive load can be improved. Additionally, they showed that the number of blinks has relation with the task difficulty: higher cognitive load leads to a decrement of the number of blinks.

Chen et. al [20] used eye-tracking technology to assess the mental effort in HCI. They conducted the experiment with basketball players who had to play games on the tablet.

During the experiment, eye activities such as blinks, pupil dilation, saccade, and fixation were recorded with the help of head mounted (mobile eye tracker) eye tracker. Authors found significant relation between the cognitive state and eye movements. They concluded that increasing number of fixations indicate more attention and decreasing number of blinks indicate higher cognitive load.

One of the most recent research on this topic conducted by Barreras [10]. The author used eye-tracking technology to assess the cognitive load in visual search. During the experiment, mobile eye tracking was used to measure eye movements such as fixation, saccade, blinks and pupil dilation. The author investigated the influence of the cognitive load to each gaze measurements and discussed how each gaze measurement can help to measure cognitive load. The results of this work show that number of fixations and saccades are increasing while increasing the cognitive load and the higher diameter of the pupils indicates higher cognitive state.

3.2 Display size in HCI

Larger displays become more affordable and gain more interest because of its capability holding more information. However, choosing the right display size to increase users' performance has been the research topic in HCI.

Karam [21] investigated the effects of the mobile device display size and the text orientation on the learning, cognitive load, and user perception. For the experiment, laptops and mobile devices were used to investigate the influence of the display size in a chemistry course. Authors concluded that learning outcome and cognitive load were unaffected by changing the display size.

Lischke et. al [22] used six different sized displays to measure the influence of the size on the users' performance. Authors used the visual search task to investigate how the task completion time depends on the large display size. Nasa TLX questionnaire was also used as a measurement. Authors concluded that larger displays do not always support users' performance.

Jain [23] investigated the influence of the display size on spatial memory. For the experiment, two different displays with different sizes (10 inches and 55 inches) and four different tasks which are related to the spatial memory were used. The author concluded that participants performed faster on the small display, however, on the large display the accuracy was significantly higher.

3.3 Summary

Through the literature research, we gained a deeper understanding of the relation between cognitive load and eye movements. Furthermore, we have learned that how the level of cognition effects the gaze measurements.

In the second part of the literature search, we have learned that choosing the right display size is one of the important topics in HCI and researchers aim to improve the interaction between the users and the systems by improving output modalities. Additionally, we saw that the size of the display has an effect on the users' performance depending on the task.

We will use the knowledge gained through the literature search for our researcher to understand the impact of the display size on the cognitive load using eye-tracking technology.

4. Research question

There are few researches conducted to investigate the influence of the display in HCI. Throughout the literature review, we saw that most works are done using traditional usability metrics such as effectiveness, efficiency and satisfaction. This was our first motivation to use cognitive load as a metric to assess the influence of the display size.

Additionally, new methods to measure cognitive load with the eye-tracking technology, especially the research done by Barreras [10] motivated us to investigate the influence of the display size on the cognitive load using eye-tracking technology.

Using the knowledge from the related work, we come up with the following research question:

How does the Display Size Influence Cognitive Load in Visual Search Tasks?

To answer our research question, we use an eye-tracking technology. We collect different gaze measurements, such as fixation, saccade, blink and pupil dilation. These measurements are collected for two different display sizes. For simplicity, we will name those displays Large and Small displays. Additionally, we use the user' performance which is the number of correct answers and NASA TLX questionnaire as a measurement. NASA TLX questionnaire is used to investigate the influence of the display size more in a subjective way.

In order to answer to our main research question, we will answer the following research sub-questions:

1. How do changes in Cognitive Load influence fixations in Large and Small displays?
2. How do changes in Cognitive Load influence saccades in Large and Small displays?
3. How do changes in Cognitive Load influence blinks in Large and Small displays?
4. How do changes in Cognitive Load influence pupil dilation in Large and Small displays?

5. How do the number of correct answers and results of NASA TLX questionnaire correlate with the results gained with the help of eye-tracking technology?

5. Pilot study

In this section, we will summarize the pilot study which is conducted before the main experiment. Haralambes et.al [24] mention that before conducting the large-scale quantitative study, many researchers conduct a pilot study to avoid time and money being wasted because of poorly designed study.

A "pilot study" is a small-scale preliminary study conducted before the main research in order to check the feasibility or to improve the design of the research [25].

We also conducted a pilot study to test the lab settings, selected task, and other necessary procedures for our design. The pilot study included all steps of the experimental design and in this chapter, we will report the first experience and improvements which had to be done to conduct the real large-scale experiment effectively.

5.1 Task

Task selection was a critical issue for our experiment. We had to choose the visual search task in such way that with its help we could increase participants' cognitive load. In visual search tasks, participant looks for a target item among other items. We call those items distractors. Our task selection based on the previous work done by Barreras [10]. We will explain the description and implementation of the task, as well as the reason behind choosing this task in section 6.1 in detail. In this section we will cover the main aspects of the task.

On the initial task, participants had to find the blue circle among distractors. Distractors can be triangle and squares in any colour, and circles in any colour except blue. Figure 2 shows the target element, which is the blue circle in our task and all possible distractors. The task has 10 runs, and each run contains 3 laps. After each lap, all elements are randomly located. The number of distractors increases linearly after each run by 10 %. This makes the task more difficult to solve after each run. However, all 3 laps within one run have the same difficulty level. Which means all laps within each run has the same number of distractors.

The reason for having 3 laps for a run is to collect more representative data for each level of difficulty.

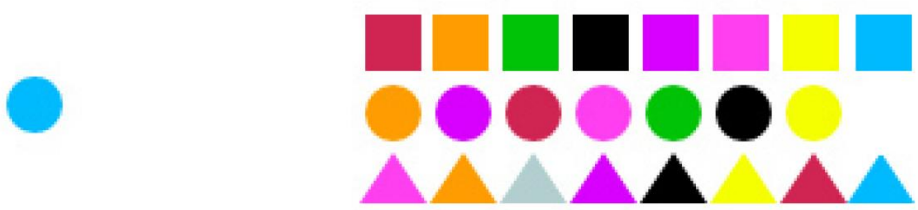


Figure 2. Target element – blue circle and all possible distractors. Figure is taken from [10]

Duration of the lap is 12 seconds, and there is a pause between each lap for 2 seconds. Duration of the whole task is 420 seconds, calculated as follows:

$$(LapDuration + PauseDuration) \cdot N°Laps \cdot N°Runs$$

As an independent variable, we chose the display size. We used 2 different sized displays. We will call these displays Large and Small displays. Display size is the condition in our experiment. Each participant performed the task on both displays.

Gaze events such as fixation, saccade, blink and pupil dilation are chosen as dependent variables. We collected these measurements using Eye-Tracker technology. With the help of dependent and independent variables, we can collect meaningful data and conduct the analysis.

5.2 Setting

We conducted the experiment in a controlled lab environment. As a large display, we used Microsoft Perceptive Pixel 55", as a small display we used Iiyama ProLite 23". It is important to keep the resolution of both screens the same, to avoid any possible influence of the different screen resolution on the cognitive load. Both displays have the same 1920x1080p resolution.

As an eye-tracker, we used SMI Eye-Tracking Glasses 2. To record the gaze events, we used Lenovo Yoga Tablet 2.

Participants performed the task while sitting on the chair. A desk placed in front of the participant and to interact with displays a mouse placed on the desk. The position of both displays, desk, and chair have been kept in the exact same position during the study to provide the same environment for all participants.

Another critical issue was to keep the luminance of the room stable. Pupil dilation is a dependent variable in our experiment. We want to study changes in the pupil dilation caused by changes in the cognitive load. However, there are several reasons for the pupillary response, and the light is one of them [15]. For this reason, we had to keep luminance stable. Daylight intensity is changing depending on the daytime and the season. In this regard, we kept the blinds of the room tightly closed and ensured that no daylight enters from the windows, and switched the lights of the lab on.

5.3 Participants

For the pilot study, we recruited 4 participants, 2 females and 2 males. They all were master students (3 participants from Computer Science, 1 participant from Political Sciences department) of the University of Konstanz. The average age of the participants was 25.25 years (min. 23 years, max. 30 years, STD=3.202). None of them had experience with eye tracking before. As they reported in Demographic Questionnaire, none of them had a vision problem (they were not wearing eye-glasses and they had no colour-blindness). However, one participant had an obvious vision problem, as he was squeezing his eyes to see the screen better. Squeezing eyes causes low tracking ratio since eye-tracker cannot detect pupils. Because of very low tracking ratio, we could not analyse the data from this participant.

5.4 Procedure

Our experiment contains several steps and sequence of these steps are important. It was useful for us to create a script and follow it for each participant to make sure, that we do not miss any step. The procedure of the experiment contains following steps:

1. Participant is welcomed, and declaration of consent introduced to the participant to read and sign.
2. Demographic questionnaire filled in by participant.
3. Eye-tracker introduced to participant and study adviser made sure that participant is comfortable with wearing it.
4. Eye-tracker calibrated, and participant started to perform the task on the first display.
5. Participant finished the task on the first display and filled in NASA TLX questionnaire.
6. Participant performed the task on the second display and filled NASA TLX questionnaire.
7. Participant took the eye-tracking glasses off and filled in post-questionnaire.

In the pilot study participant did not receive compensation for their help.

Detailed description of the procedure as well as the documents used during the experiment will be explained in the following chapter.

5.5 Lessons learned

As we mentioned earlier the idea behind conducting a pilot study is to detect shortcomings of the research design and improve them for the large-scale experiment to avoid waste of effort, time and money. We also faced many difficulties and pilot study helped us to improve our design for better results. In this section, we will discuss these shortcomings.

Selection of the chair

The first problem in our design was the selection of the chair, where the participant is sitting while performing the task. We placed displays in determined places and adjusted the height of displays according to the average height of participants and we could not change the height of the displays for individual participants. We also wanted to ensure that all participants view

the display from the same visual angle, however, this was difficult to maintain because of the significant differences in the participants' height (We will discuss experimental set up in chapter 6.3 in details). Therefore, for the large-scale experiment, we decided to use height adjustable chair.

Explanation of the task

The second problem was the explanation of the task. The task includes some necessary steps, which participants should enact. These steps are essential later in our analysis. Therefore, we had to make sure that all participants perform the task in a correct way. When participants enter the lab, we introduced them the Welcome letter and explained the proper ways to perform the task in a written form. Later, these rules explained to the participants verbally as well and asked if they have any questions regarding the task. Even if participants notified that all rules are clear, and they are ready to perform the task, we could observe that the task is not totally clear to them. In some cases, it was clear from participants' facial expressions that they do not know what to do now: if the run is finished and they can already start the next run, or they should wait until the next run starts even though they found the right answer. Additionally, we could inspect the participants' gaze movements while they perform the task and we could observe that in some cases, their gazes are not fixated on the right area, although we explained it before starting the task. For this reason, we decided to explain the task practically as well, viz. to perform few runs of the task until participants understand every detail.

Calibration process

The third challenge in our design was the calibration process of eye-tracker. This process is important for the data collection. This process is usually done only one time for each participant. However, this may not be enough for eye-tracker to detect pupils. In some cases, eye-tracker needs more time to detect pupils properly or study adviser must calibrate the device more than once. The long calibration process is highly dependent on the physiognomy

of eyes. This process is time-consuming and sometimes can be frustrating for the participant, however, poorly calibrated eye-tracker causes offsets during the recording.

Duration of the task

The last challenge we faced in our initial study was the duration of the task. In the pilot study duration of each lap was 12 seconds. In earlier runs, where the difficulty of the task is not high, 12 seconds was fairly enough to find the right answer. However, starting from the 4th run most participants had difficulties to find the right answer on time and they were complaining about very limited time. For this reason, we decided to increase the duration of the lap from 12 to 15 seconds, and reduce the additional stress caused by limited time. Likewise, we increased the duration of the pause between laps from 2 to 3 seconds. In some cases, 2 seconds was not enough for participants to notice that the task is over, they should stop looking for the element and wait until the next lap starts.

The pilot study was helpful for us to consider aforementioned issues and improve them. These challenges helped us to redesign our study and conduct the large-scale experiment effectively.

6. Experimental Design

In this chapter, we will discuss the large-scale experiment in detail. Based on the pilot study, we redesigned our study and took relevant actions to improve shortcomings for better data collection process.

This chapter covers detailed description of the task, apparatus which had been used, participant recruitment and whole procedure to conduct the study.

6.1 Task description and implementation

Visual search tasks usually involve active scanning of the stimuli [1] to find the target element among other elements (distractors) [26]. Visual search tasks are part of our everyday life. We are often looking for relevant items: book in the library, food in the grocery store etc. However, there are special search tasks which are designed to conduct experiments. These tasks are mainly divided into two categories: feature search task (disjunction) and conjunction search task. In feature search tasks, distractors differ from the target element only by one feature, such as colour, shape, size, orientation. However, in conjunction search tasks, distractors differ from target element by more than one feature [27]. Figure 3 illustrates examples for both search categories.

[1] A stimulus is any visual content, which participant is looking at during the eye-tracking experiment.

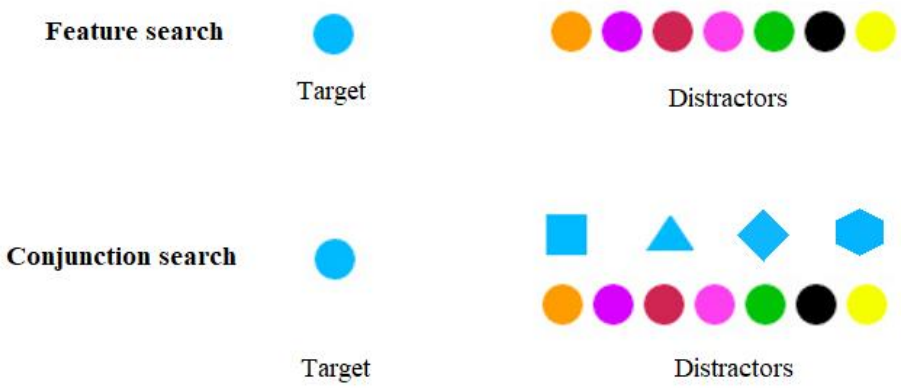


Figure 3. In feature search task, distractors are circles only and they differ from the target by color. In conjunction task distractors are circles in different colours and shapes in color blue, they differ from the target by color and shape respectively.

Selection of the task to use in our experiment is based on the previous study conducted by Barreras [28]. She proposed a new way to measure cognitive load in the visual search task. In her study, she used eye-tracking technology to measure gaze events. As a visual search task, she used 3 different tasks:

- Condition colour: distractors differ from the target by their colour (feature search task)
- Condition shape: distractors differ from the target by their shape (feature search task)
- Condition colour & shape: distractors differ from the target by their shape and colour (conjunction search task)

After analysing collected data, Barreras showed that 3rd condition (colour&shape) is more suitable to measure cognitive load with eye-tracking technology, as the gaze events were changing significantly while increasing difficulty of the task, which leads to an increment on the cognitive load. In this regard, we decided to use the search task where the participant is looking for a blue circle among triangles and square in any colour, and a circle in any colour except blue (See Figure 2).

The visual search task is implemented by Barreras[10]. To implementation the task JavaScript[2] programming language, and for the interface HTML5[3], and CSS3[4] have been used. The task has two main screens: log screen and a task screen.

Settings of the task can be configured on the log screen (see Figure 4). Before participant starts the task, study adviser can input the id number of participant, the task mode – colour&shape in our design, the number of runs, and the number of laps. After inputting all necessary data, participants can start the task by clicking on the *"Start Experiment"* button located on the bottom-left corner of the screen. Once participants start the experiment the task screen appears. Task screen contains distractors and the target element. When the participant finds the target element and clicks on it beep sound generated to confirm that the click was on the target element.

Figure 4. The log screen of the task with the configuration panel.

[2] https://www.javascript.com
[3] https://www.w3.org/html
[4] https://www.w3.org/Style/CSS/Overview.en.html

The size of elements set to 20px. Elements placed on a grid with 75 columns and 36 rows. The maximum number of elements can fit to screen is $75 \cdot 36 = 2700$.

The number of elements is increasing linearly after each run. We set the number of runs to 10, which means the number of distractors increases by 10 % after each run, calculated as follows:

$$Items\ increment = \frac{Max\ N°items}{runs} = \frac{2700}{10} = 270$$

Only in the first run this number is 269, to reserve a place for the target element. In the last run, the number of elements reaches its maximum and covers the whole screen. Figure 5 illustrates various stages of the task.

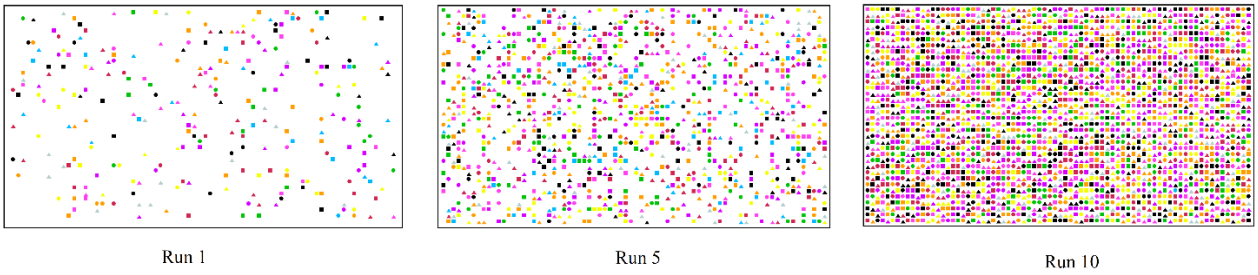


Figure 5. The Task screen on 3 different stages: Run 1, Run 5 and Run 10.

In a quantitative experiment, it is important to collect enough sample to make the data more representative. For this reason, each run of the task contains 3 laps. All 3 laps within one run have the same number of distractors, as shown in Figure 6.

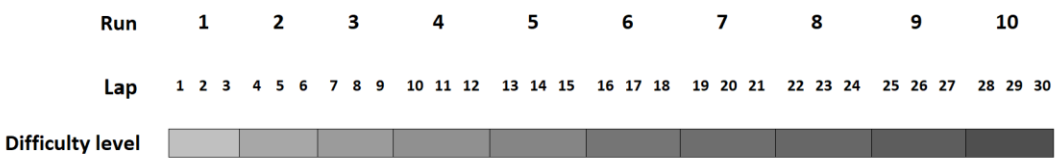


Figure 6. Structure of the runs and laps with difficulty level.

As we already discussed in the pilot study, initially the duration of the task was 12 seconds and finding the target item within 12 seconds was very challenging for participants starting from the 4th run. For the main experiment, we decided to increase this duration from 12 to 15 seconds to provide participants with more time. After 15 seconds the lap is over, and it is not possible to click on the target element anymore. On this time the resting screen appears (see Figure 7). Purpose of the resting screen is to redirect gaze movements to the middle of the screen and make sure that every participant starts the next lap from the same position. However, the main objective of the resting screen is to collect gaze movements while the participant is not in active search. We will discuss details of the gaze data collected during the task and during the resting state in the next chapter.

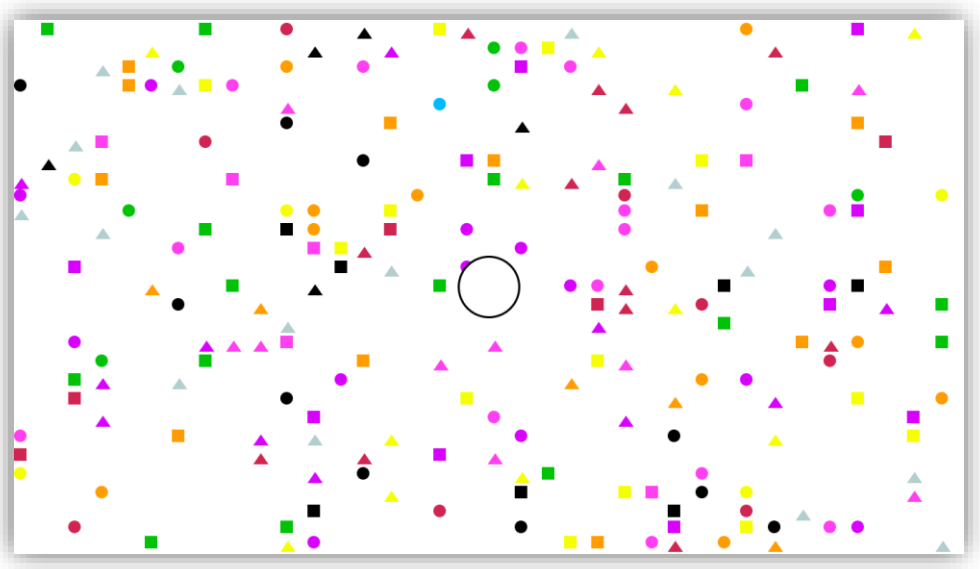


Figure 7. Resting screen with the circle in the middle to redirect participants' gaze movements.

The resting screen visually does not differ from the previous task screen. To notify the participant that the task is over, a small circle appears in the middle of the screen. At this point, the participant should stop looking for the element, bring the mouse pointer inside of the circle and concentrate to the middle of the screen. In the pilot study, we observed that in some cases, 2 seconds were not enough for participants to recognize the circle and stop looking for the target element. For this reason, we increased the duration of the resting state as well. The idea behind keeping the stimuli in the resting state the same as in the task

duration (active state) is to keep luminance the same. Changing the whole stimuli for the resting state would easy for participants to recognize, however, this might change the intensity of the light coming from the screen and lead pupillary response.

Additionally, if participants found the right answer within 15 seconds, we asked them to bring the mouse pointer to the middle of the screen and concentrate to the middle of the screen until the lap is over and resting state appears.

In the end of the task log file is created. The log file contains the necessary information for the analysis of the collected data. The log file contains the following information:

- Participant ID
- Mode (colour&shape in our experiment)
- Number of runs and laps
- Number of elements added after each run (270 in our experiment)
- Target element (blue circle in our experiment)
- Start and end time of the task in HH:MM:SS:MS format
- Number of wrong clicks
- Current run
- Current lap
- Start and end time of the current lap in HH:MM:SS:MS format
- Time of the click in the HH:MM:SS:MS format

This information will help us to analyse the data efficiently.

6.2 Independent and dependent variables

There are two types of variables in the experimental analysis: *independent* and *dependent* variables. Independent variables – variables that are controlled by the examiner. Dependent variables – variables that are measured during the experiment. In the analysis, researchers assume that independent variables have an effect on the dependent variables [29].

To answer our research question with the help of eye-tracking technology, we have chosen following gaze events as the dependent variables:

- Fixation
- Saccade
- Blink
- Pupil dilation

As we already discussed in the literature review section, there is a relation between these eye-tracking measurements and the level of cognitive load and these measurements will help us to answer to our research question in an objective way. Additionally, we included two more dependent variables to assess the cognitive load. These variables are the performance which is the number of right answers given on each display and the NASA TLX[5] questionnaire. We used NASA TLX questionnaire to assess the influence of the display size on the cognitive load more in subjective way.

NASA TLX is a rating procedure that includes six subjects to rate overall workload score [30]. These subjects are the followings:

- **Mental demand**: How much mental and perceptual demand, such as thinking, deciding, remembering, searching, etc. required for the participant to solve the task?
- **Physical demand**: How much physical activity, such as pushing, pulling, turning etc. required for the participant to solve the task?
- **Temporal demand**: How limited was the given time to solve the task on time?
- **Performance**: How successful did participant complete the task?
- **Effort**: How mentally and physically challenging was the task to perform?
- **Frustration**: How insecure, irritated, annoyed etc. did participant feel while performing the task?

As we already discussed in the research question, we want to study the impact of the display size on the cognitive load. For this reason, we used the display size as an independent variable

[5] Nasa Task Load Index. https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20000021488.pdf

and we assume that the display size (independent variable) has an effect on the cognitive load which is measured with the help of gaze events (dependent variables).

For our experiment, we used 2 different sized displays (large and small displays). All participants performed the task in both displays. Selection of the displays will be discussed in the following chapter.

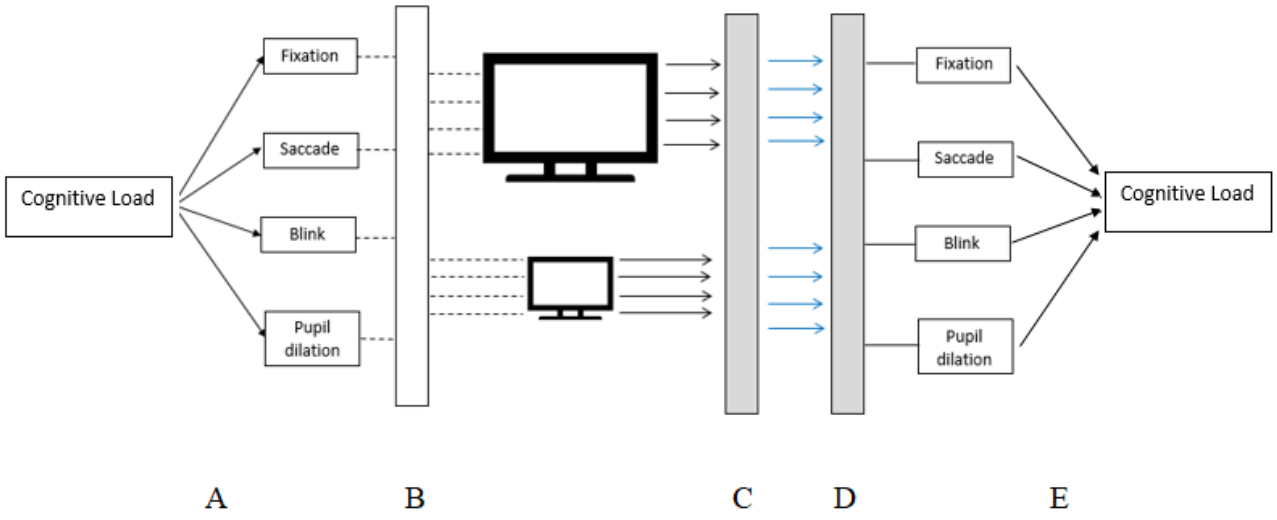


Figure 8. Structure of the necessary steps to answer our research question.

Figure 8 illustrates our dependent and independent variables on the different stages of the experiment.

- **A**. There is a relation between cognitive load and gaze events as we have seen in the literature review.
- **B**. While performing the visual search task, we collect measurements of four gaze events, on both, large and small displays.
- **C**. 1st phase of analysis: we analyse these gaze event for each screen. Changes on the gaze measurements depending on the level of cognition are taken into account. To be specific, we analyse how the gaze events change during the active search (higher cognitive load) and during the resting state (lower cognitive load). We investigate if

the level of cognition influenced the eye-tracking measurements in the same way for both displays.

- **D.** 2nd phase of analysis. We make a comparison between gaze events collected on large and small displays to study how do gaze events change between displays.
- **E.** With the knowledge gained in **C** (how the cognitive load influence gaze measurements) and **D** (how gaze measurements change depending on the display size), we discussed how the display size affect the level of cognition.

6.3 Apparatus

We conducted the experiment in the controlled lab environment. In this kind of environment, we could minimize the factors which might influence the dependent variables. For instance, the non-stable intensity of the light may cause to the pupillary response, for this reason, we had to keep the luminance of the room stable by closing the blinds and switching the lights of the lab on. To avoid distraction, we also closed the door of the lab and made sure that no one can enter the lab while the participant is performing the task.

For the display choice, we decided to select devices that are used in our real life. The large screen which is used in public places, such as airports, train stations to check the timetable and small screen which is mostly used for the desktop computers. As a large display we have selected Microsoft Perceptive Pixel 55"[6] and as a small display Iiyama ProLite 23"[7].

Table 1. Technical characteristics of selected displays

Name	Resolution	Display Area	Aspect ratio	Luminance
Perceptive Pixel 55"	1920×1080	47.6 × 26.8 inches 121 × 68.1 cm	1.78	400 nits
Iiyama ProLite 23"	1920×1080	20 × 11.3 inches 51 × 28.7 cm	1.78	225 nits

[6] http://www.perceptivepixel.com/
[7] https://www.iiyama-monitors.co.uk/categories/monitors/iiyama-23-and-24-inch-Monitor.aspx

We want to know how the size of the display affects the cognitive load, for this reason, we have to keep other characteristics of the displays the same except the size. As described in table 1, displays have the same resolution and aspect ratio to avoid any possible influence of these factors to the level of cognition. However, the luminance level of the displays differs 400 nits for the large, 225 nits for the small display. Microsoft Perceptive Pixel has the adjustable brightness from 1 to 32 [31]. For this reason, we decreased the brightness of the large display from the level 32 to the level 16 during our experiment, to ensure that both displays have the same luminance level.

As we discussed in the pilot study, for the main experiment we decided to use height adjustable chair to ensure that all participants' gazes centered to the middle of the displays while sitting and they view the display from the same visual angle.

The height of the screens are adjusted as follows: We defined that the average distance from the floor to the eye level (while the participant is sitting) is 130 cm. This means, the distance from the floor to the middle of both displays is 130 cm as well. The height of the large screen is 80 cm and the small screen is 33 cm. The distance from the floor to the top of the screens have been calculated as follows:

For the large screen: $130 + \frac{80}{2} = 170\ cm$

For the small screen: $130 + \frac{33}{2} = 146.5\ cm$

We kept these adjustments the same for all participants. Additionally, we could change the height of the chair to ensure that participants' gazes centered to the middle of the displays.

To maintain the same visual angle for both screens we have done some calculation as shown in Figure 9. Our aim was to keep the visual angle for small display $\angle\beta_1$ the same as the visual angle for the large display $\angle\beta_2$.

First, we defined the distance to view the small display. We decided that 80 cm between the participant and the small display is comfortable to perform the task ($S = 80\ cm$). After this, we could calculate the visual angle for the small display, $\angle\beta_1$. For this, we used the display area of the screen (Table 1), which is $h_S = 28.7\ cm$ and calculated $\angle\beta_1$ as follows:

$$\tan \beta_1 = \frac{S}{\frac{h_S}{2}} = \frac{80}{\frac{28.7}{2}} = \frac{80}{14.35} \approx 5.57$$

$$\beta_1 = \arctan 5.57 \approx 80°$$

The visual angle for the small screen is $\angle\beta_1$ =80°, which means the visual angle for the large screen $\angle\beta_2$ must be 80° as well. Using the height of the large display, $h_L = 68.1\ cm$, from Table 1, we can calculate the distance from the large screen to the participant (L):

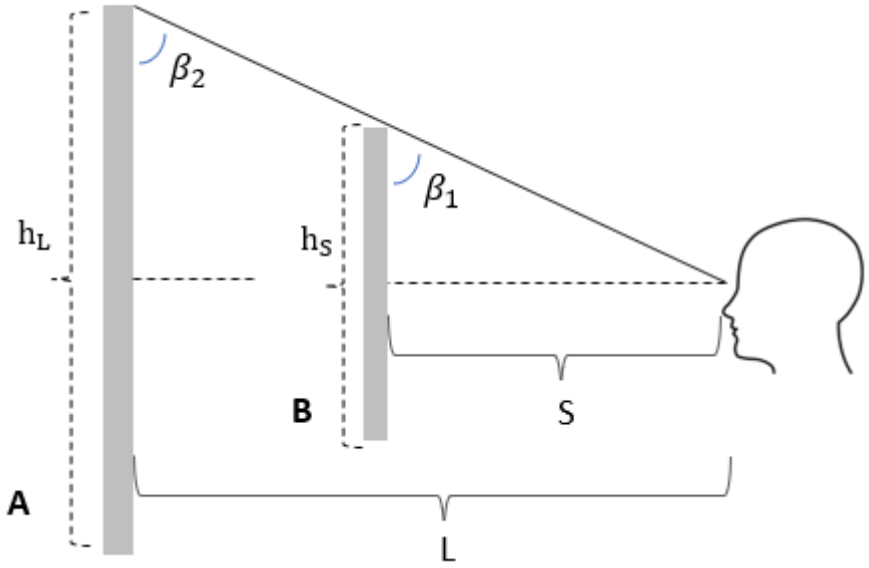


Figure 9. Placement of the displays. A - Large display, B - Small display

$$L = \frac{h_L}{2} \cdot \tan \beta_2 = \frac{68.1}{2} \cdot 5.57 \approx 190\ cm$$

Once we have the height of each display and the distance from the participant to each display, we can set up the lab (see Figure 10).

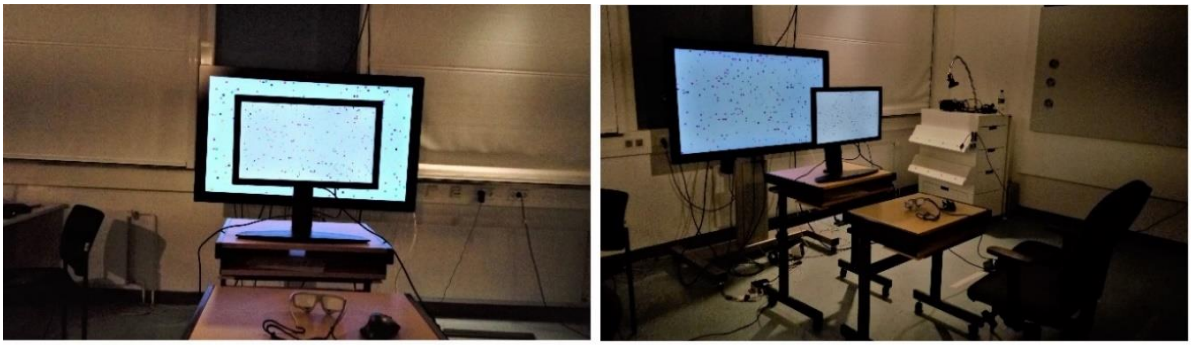


Figure 10. Replacement of the displays. For demonstration purposes, both displays are on. Eye-tracker and the mouse placed on the desk.

For the experiment, we used SMI Eye-Tracking Glasses 2[8]. These are remote eye-tracking glasses, which is an advantage for our experiment: during the recording, participants can move their head without stopping the recording process. This is especially important while filling the NASA TLX questionnaire (details are explained in the following section). Eye-tracking glasses were connected to the smart recorder, Lenovo Yoga Tablet 2[9]. SMI Iview ETG 3.7 software was installed on the smart recorder. With the help of this software, study adviser can configure eye-tracker.

Eye-tracking glasses are operating at 60 HZ and recording one video and one audio file for each participant. After finishing recording process, collected data transferred to the laptop to process the data, with the help of Begaze 3.7 software (Details are discussed in the following chapter).

6.4 Participants

Participants were recruited using Doodle poll[10] online scheduling. We prepared flyers with basic information related to the experiment, such as name, date, location, doodle link and the amount of compensation for the help.

[8] https://www.smivision.com/
[9] https://www3.lenovo.com/us/en/tablets/lenovo/yoga-tablet-series/yoga-tablet-2-win-10/
[10] https://doodle.com/de/

We recruited those participants who do not wear eyeglasses. Eye-tracking glasses does not work properly when it is worn over the eyeglasses. However, using eye-tracking glasses with contact lenses were acceptable for the experiment. Another vision problem which was not acceptable for the experiment was the colour blindness. Our task involves many colours and we wanted participants to clearly distinguish all colours.

In the eye-tracking experiment, it is important to conduct the experiment with a higher number of participants than planned for analysis. Our aim was to analyse the data from at least 30 participants. For this reason, we conducted the experiment with 35 participants to ensure that in case of erroneous data resulted by device failure or significant outliers, we can exclude those data from the analysis and still have enough data to analyse.

18 females and 17 male participants took a part in the experiment. We did not control the participant selection by their gender, however, it is an advantage of the experiment to have an almost equal number of female and male participants. The average age of the participants was 24.8 (min. age 19, max. age 31, STD = 3.21) and they all were students from Konstanz University (12 participants bachelor, 18 participants master, 5 participants Ph.D.). Educational background of the participants was diverse, such as Computer Science. Life Sciences, Politics and Public Administration, Literature, Linguistics, Sociology, Psychology etc. A wide range of backgrounds of participants makes out study more representative. Average height of the participants was 171.12 cm (min. 153cm, max.188cm, STD = 10.31). 6 out of 35 participants were using contact lenses during the experiment. 17 participants had no experience with eye-tracker before, 8 participants used only once, 10 participants involved in an eye-tracking experiment more than once.

We opted to use the within-subject design for our experiment. In a within-subject design, all participants are assigned to both conditions. The most important advantage of this design is that we do not need a large number of participants to collect enough data [32]. All 35 participants performed the task on both displays. However, there is also disadvantages of this design: performing the task in one condition may affect the performance in other condition. This is called a learning effect, which can cause a bias in the research. To prevent learning effect, we used a Latin square to counterbalance the conditions: if the current participant starts performing the task on the large display and continues with the small display, then the

next participant starts performing the task on the small display and continues with the large display.

6.5 Procedure

It is important to have a formal procedure during the experiment. It helps researchers to follow the same instruction for every participant, and make sure that all participants have the same experience and knowledge about the study [33]. We also created a script and included all the necessary steps to ensure that we conduct the experiment in exactly the same for all participants.

According to Hornbæk [34], it is very important to treat participants with respect, consider their feelings, needs and well-being. He also described key principals of ethical issues. This was very important for our experiment as well, and we will discuss these key principals in the frame of our study:

- ***Voluntary participation and informed consent***. Participants should decide if they want to participate or not, and they can change their decision anytime. The experimenter has to give all necessary information before starting the experiment, so participants can still stop the experiment if they do not want to continue. In our experiment, we gave all necessary information about the task in the Welcome letter (see Appendix A), additionally we mentioned that we will collect gaze measurements including the pupil size. However, in the Letter of Consent (see Appendix A), we asked participants' understanding that before performing the task and collecting measurements we cannot give detailed information about our purpose and expectations. However, at the end of the experiment, we were happy to explain our study in detail and answer all questions related to study. Explaining the study before collecting gaze movements would lead to a biasing in the results since participants could imitate their voluntary gaze movements. At the end of the experiment, participants still had rights to ask us to remove their data from the experiment.
- ***Protection from harm***. Participants in HCI experiment should be protected from harm. Like most electronic devices, eye-tracking glasses produce some heat after long

usage. We made sure that after the experiment with each participant, we switch the eye-tracker off and let it cool down. Before wearing eye-tracker, we informed participant about it. We explained that after some time, it is normal for glasses to heat and they can take the glasses off if they do not feel comfortable. During the whole experiment, we did not have any problem regarding this and all participants were feeling comfortable with the temperature of the device.

- ***Privacy***. Collected data must be kept anonymous. During the publishing, the experiment participants should be anonymous as much as possible. Researchers have to be careful while using pictures and videos of participants where the faces are recognisable, or including names should be avoided. In the letter of consent, we declared that the collected data will be anonymized and will completely be used for research purposes within the framework of the study.
- ***Contact to researchers***. Participants should be able to contact to study adviser after the experiment for any reason. We were sending confirmation emails to each participant while inviting them to the experiment. And ensured that all participants have our contact Email address.

In the following, the whole procedure of the experiment is explained.

The study adviser prepares the room by closing the blind and putting “do not disturb” sign on the door. All necessary documents for the experiment are prepared. The display which the participant will perform the task first with switched on. For simplicity, we implied that the participants with odd id number will start the task by performing it on the large display, otherwise on the small display. Eye-tracker and the mouse placed on the desk.

When the participant enters the lab, examiner welcomes the participant and introduces himself. Examiner starts a small talk to create a friendly environment and reduce the stress. After the participant sits on the chair, examiner gives brief information about the study and introduces the Welcome letter. If the participant has no more questions regarding the procedure, examiner hands out the Letter of Consent (see Appendix A) to the participant to read and sign. Additionally, the participant has to fill demographic questionnaire (see Appendix A) before starting the task.

Examiner introduces both displays and eye-tracker to the participants which they will use during the experiment. As we discussed in the pilot study, we decided to explain the task to the participants not only in written form and orally, as well as practically. We concluded that the most effective way of explaining the task is performing the task for participants and explaining all details in practice. Examiner starts performing the task for two to three laps and at the same time explaining all details which has to be done correctly while recording process, such as concentrating to the middle of the screen after finding the target element or stop looking for the element after recognizing the circle in the middle of the screen and locate the mouse pointer inside of it. As long all details are explained to the participant, we finish the training procedure.

Once the participant has no more question about the task we can proceed to the next step: wearing and calibrating the eye-tracker. With the help of examiner, participant wears the eye-tracker. Examiner makes sure that participant feels comfortable with eye-tracking glasses on. Now the eye-tracker can be calibrated. Calibration is the process where the eye-tracker detects the pupils and reflects the position of the gaze the same as in the real world. As we already discussed in the pilot study, the calibration process may take longer depending on the physiognomy of eyes and may be frustrating for the participant. However, we had to spend enough time on this process to calibrate the device as accurate as possible and ensure that we will gather correct data. We chose 3-point calibration, which means participants will have to concentrate on three different points while we calibrate the eye-tracker (see Figure 11).

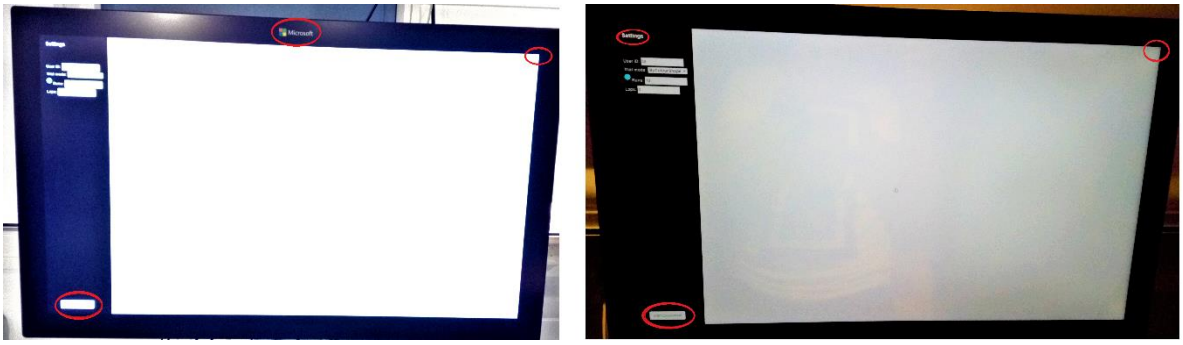


Figure 11. Calibration points for the Large (left) and the Small (right) displays. For the large display "Start the experiment" button, the Microsoft sign and the top right corner of the display are selected as calibration points. For the small display "Start the experiment" button, the word "Settings" on the top left corner of the display and the top right corner of the display are selected as calibration points.

After the successful calibration process, examiner starts the recording process. From now eye-tracker is gathering all the necessary gaze measurements. Participants are informed that they can start the task whenever they feel ready.

After the completing the task for one condition, participant fills the NASA TLX questionnaire (see Appendix A). We asked participants not to take the eye-tracker off. Taking the glasses off would prolong the experiment, since we would have to stop the recording after performing the task on the first display and calibrate the eye-trackers again for the second display. While participant fills the questionnaire with eye-tracking glasses on, examiner prepares the other display (see Figure 12).

Now participant can start performing the task on the other display. After completing the task on the second display, examiner stops the recording process. The participant takes the glasses off and fills second NASA TLX questionnaire.

Figure 12. Left: Performing the task on the large display while the small display placed aside. Right: Performing the task on the small display while the large display is switched off.

After finishing the task on both displays, participants fill one last questionnaire related to their experience on both displays and they can also make their comments verbally.

In the end, the examiner asks participants some open-ended questions (see Appendix A).

Duration of the whole experiment with each participant was approximately 45 to 55 minutes.

Finally, we thanked the participants for their help and they received monetary compensation for their time.

7. Analysis

In this section, we will summarize the data preparation and the data analysis process. After collecting a large amount of quantitative data, analysis part is one of the most time-consuming section of the research.

In the data preparation section, we will discuss every step of transforming the raw data and make it ready for analysis. After finishing data preparation part, we will conduct several statistical analyses to answer our research question.

7.1 Data preparation

After recording the data with the help of SMI Iview ETG 3.7 software, data has to be extracted and all necessary measurements need to be obtained. Iview ETG software records one audio and one video file for each participant. To extract these data, SMI provides another software called BeGaze (Behavioural and Gaze Analysis) which is fully integrated with Iview ETG.

BeGaze software provides the list of all recorded data for each participant and basic information for each data, such as length, recording time, number of samples (including the number of each gaze event gathered during the recording process) and most importantly the tracking ratio (see Figure 13). Tracking ratio indicates the quality of the recorded data, the higher the information ratio the better quality. In our experiment the average tracking ratio was 98.70 % (min. 91.40%, max. 100%, STD = 1.77).

Furthermore, the software allows examining the recorded file frame by frame. It is important to check if the eye-tracker had offsets during the recording process. If there is an offset, the with the help of offset correction function the problem can be solved.

The metric export functionality of the BeGaze software exports the necessary information selected for each participant. To analyse the collected data, we exported following data for each participant:

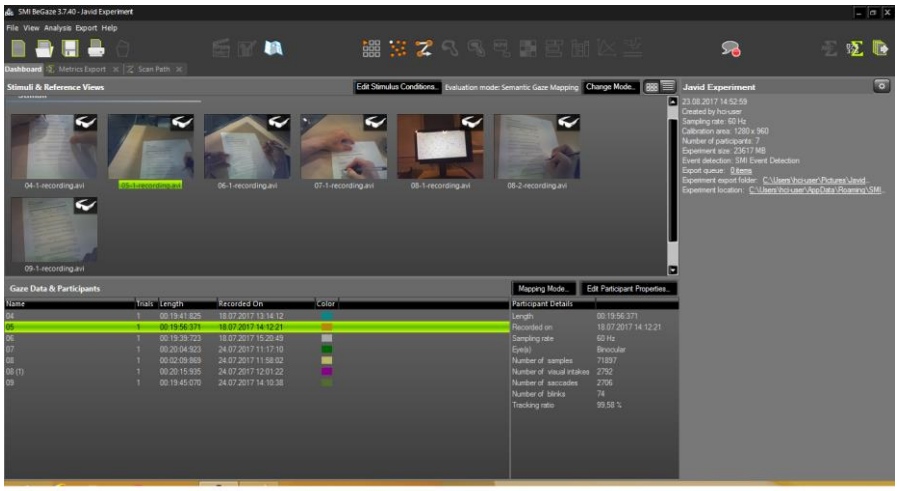


Figure 13. Main screen of the BeGaze software. It provides the list of all recorded files and basic information for each file.

- **Participant**: the id number of the participant.
- **Category**: the name of the gaze event, such as fixation (described as visual intake in the software), saccade and blink.
- **Event Start Video Time [ms]:** starting time of the gaze event. The time started from the beginning of the recording, e.g.: "event start video time" of the visual intake is 00:00:00:083 means that the event was measured 85 milliseconds after the starting recording (see Figure 14).
- **Event End Video Time [ms]**: ending time of the gaze event.
- **Visual Intake Average pupil diameter [mm]**. The size of the pupil recorded during the fixation (visual intake) event.

Participant	Category	Event Start Video Time [ms]	Event End Video Time [ms]	Visual Intake Average Pupil Diameter [mm]
12	Visual Intake	00:00:00:083	00:00:00:208	2.8
12	Blink	00:00:00:208	00:00:00:500	-
12	Visual Intake	00:00:00:500	00:00:00:667	2.9
12	Saccade	00:00:00:667	00:00:00:750	-
12	Visual Intake	00:00:00:750	00:00:01:042	2.9
12	Blink	00:00:01:042	00:00:01:458	-
12	Visual Intake	00:00:01:458	00:00:01:625	2.8
12	Saccade	00:00:01:625	00:00:01:708	-
12	Visual Intake	00:00:01:708	00:00:01:833	2.9
12	Blink	00:00:01:833	00:00:02:751	-
12	Visual Intake	00:00:02:751	00:00:03:043	2.8
12	Saccade	00:00:03:043	00:00:03:168	-
12	Visual Intake	00:00:03:168	00:00:03:293	2.7

Figure 14. Extracted raw data for participant id=12.

The log file, obtained at the end of the task contains start and end time of each lap in real time. Our aim in the data preparation process is to synchronise the raw eye-tracking data and the log file. For this, we did several steps which we will explain in this section.

For each participant, we collected two log files (one for each condition) and one gaze data. To automatize the process, we used Java programming language for data preparation. Java project with all used classes is submitted in electron version on the submission of this thesis.

To ensure that we do not miss any gaze event during the participant performs the task, we started the recording process before the participants start the task, and finished the recording after participant finishes the task.

First, we changed the timestamps in the log files. As in the eye-tracking data, we converted all timestamps of the lap to the video starting time. To do this, we found the differences between the real time in the log file and the real time of the beginning of the recording process and assigned the new values to the corresponding timestamp in the log file, e.g.: if the starting time of the lap is 10:06:30:881 and the starting time of the recording process is 10:06:05:059, then the new value for starting time of the lap is $10:06:30:881 - 10:06:05:059 = 00:00:25:822$ .This is done with the help of *ChangeTime.java* class.

We want to analyse the eye-tracking measurements recorded while the participant performs the task and while the participant is in the resting state. In this regard, we defined **OnRun** and **OffRun** time intervals:

When the participant found the target element and clicked on it:

- **OnRun**: duration which starts from the beginning of the lap and lasts until the participant clicks on the target element.
- **OffRun:** duration which starts from the finishing point of the **OnRun** interval and lasts until the end of the lap.

When the participant did not find the target element:

- **OnRun:** duration which starts from the beginning of the lap and lasts 15 seconds (duration of one lap).
- **OffRun:** duration which starts from the finishing point of the **OnRun** interval and lasts 3 seconds (duration of one pause).

Defining intervals for each log file is done with the help of *DefineIntervals.java* class. Figure15 summarises the transformation of the log file.

Run	Lap	Start Lap	End Lap	Clicks
1	1	10:06:30:881	10:06:45:882	10:06:35:156
1	2	10:06:45:882	10:07:03:883	10:06:55:324
1	3	10:07:03:883	10:07:21:885	10:07:10:012
2	1	10:07:21:885	10:07:39:886	10:07:29:457
2	2	10:07:39:886	10:07:57:886	10:07:51:521
2	3	10:07:57:886	10:08:15:887	10:08:08:704
3	1	10:08:15:887	10:08:33:888	10:08:25:703
3	2	10:08:33:888	10:08:51:890	10:08:43:687
3	3	10:08:51:890	10:09:09:891	-
4	1	10:09:09:892	10:09:27:893	-
4	2	10:09:27:893	10:09:45:893	10:09:37:963

a)

Run	Lap	Start Lap	End Lap	Clicks
1	1	00:00:25:822	00:00:40:823	00:00:30:097
1	2	00:00:40:823	00:00:58:824	00:00:50:265
1	3	00:00:58:824	00:01:16:826	00:01:04:953
2	1	00:01:16:826	00:01:34:827	00:01:24:398
2	2	00:01:34:827	00:01:52:827	00:01:46:462
2	3	00:01:52:827	00:02:10:828	00:02:03:645
3	1	00:02:10:828	00:02:28:829	00:02:20:644
3	2	00:02:28:829	00:02:46:831	00:02:38:628
3	3	00:02:46:831	00:03:04:832	-
4	1	00:03:04:833	00:03:22:834	-
4	2	00:03:22:834	00:03:40:834	00:03:32:904

b)

StartTimeOn	EndTimeOn	StartTimeOff	EndTimeOff
00:00:25:822	00:00:30:097	00:00:30:097	00:00:43:823
00:00:43:823	00:00:50:265	00:00:50:265	00:01:01:824
00:01:01:824	00:01:04:953	00:01:04:953	00:01:19:825
00:01:19:826	00:01:24:398	00:01:24:398	00:01:37:827
00:01:37:827	00:01:46:462	00:01:46:462	00:01:55:828
00:01:55:827	00:02:03:645	00:02:03:645	00:02:13:828
00:02:13:828	00:02:20:644	00:02:20:644	00:02:31:829
00:02:31:829	00:02:38:628	00:02:38:628	00:02:49:830
00:02:49:831	00:03:04:832	00:03:04:832	00:03:07:832
00:03:07:833	00:03:22:834	00:03:22:834	00:03:25:834

c)

Figure 15. a) The original log file created in the end of the task. b) log file with converted timestamps, from the real time to the video starting time. c) Log file with the timestamps for the beginning and the end of the OnRun and OffRun periods.

After transforming the log files, with the help of *GazeData* class, we synchronised the log files with the eye-tracking data. The program reads the transformed log file, and the corresponding file with gaze measurements and creates another file containing the average number of gaze events per second for each run on the OnRun and OffRun period. Gaze events which are taking place between two intervals, e.g.: if the saccade starts on

the OnRun period and finishes on the OffRun period, are excluded from the analysis, (see Figure 16).

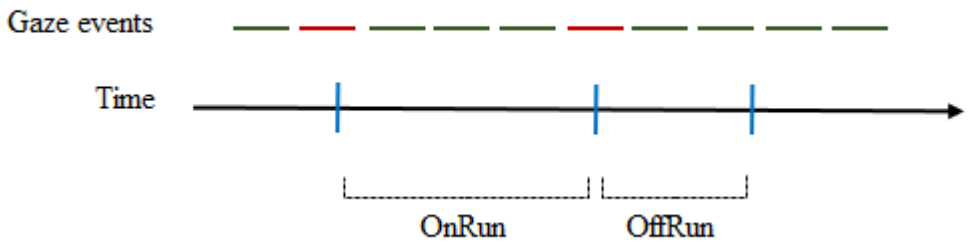


Figure 16. Red colored gaze events are excluded from the analysis since they cross the border between OnRun and OffRun.

After the synchronization process, we obtain two files for each participant (one file for the large, one file for the small display). Each file contains the average number of gaze events per second for each category (fixation, saccade, blink, pupil diameter) during on and off run period (see Figure 17).

Run Nr.	BlinksON	BlinksOFF	FixationsON	FixationsOFF	SaccadesON	SaccadesOFF	Average pupil	Average pupil diameterOFF
1	0.044441	0.255105	2.3552291	1.533753	1.9830381	1.1675371	2.368083	2.2881904
2	0.075999	0.088777	1.7153782	2.070753	1.7994894	1.9567553	2.2794445	2.2625
3	0	0	2.1243904	2.2096121	2.0434759	2.0088406	2.276326	2.2873447
4	0	0.156546	2.920356	2.3712623	2.7316685	2.0100546	2.301623	2.2731783
5	0.066662	0.222222	2.6442769	2.1111112	2.599837	1.6664816	2.281593	2.370476
6	0.022221	0.111483	2.323928	1.6016351	2.1035707	1.26793	2.3086843	2.325
7	0	0.132837	2.3886068	1.3039378	2.2552855	1.0382642	2.2511969	2.2607143
8	0.022221	0.150765	2.661505	2.6105163	2.5031414	1.9756522	2.263189	2.4068782
9	0.044441	0	2.733209	2.5552223	2.5554414	1.9998149	2.256724	2.226389
10	0.044441	0.222222	2.5109437	1.4444443	2.3553984	1.4444443	2.235335	2.218889

Figure 17. An example of the file for participant 12 in condition Large display.

Finally, to make the data ready for statistical analysis, we rearranged the data with the help of *NewArrange.java* class. After this step, we obtained eight files: one file per gaze event for both displays. Each file contains the average number of the gaze event per second for each participant per run (on and off duration). Additionally, we added one column for each run called $R_i diff$ which is the difference between the values of the on and off periods (see Figure 18).

Part.	R1ON	R1OFF	R1diff	R2ON	R2OFF	R2diff	R3ON	R3OFF	R3diff
P1	2.886467	2.048604	0.837864	2.827043	2.865809	-0.03877	2.96493	2.374711	0.590219
P2	2.807652	2.699553	0.108098	3.079728	2.227276	0.852453	3.464515	2.217529	1.246985
P3	1.883974	1.109	0.774975	2.255869	1.208992	1.046878	2.702215	2.14681	0.555405
P4	2.781822	1.146734	1.635088	3.454938	0.78163	2.673308	2.537786	1.682046	0.85574
P5	2.84358	1.698523	1.145057	2.668968	1.673005	0.995963	3.004034	2.177883	0.826151
P6	2.482285	0.656683	1.825602	2.963063	0.731493	2.23157	2.76825	0.820982	1.947267
P7	2.303685	2.504067	-0.20038	2.001614	1.910688	0.090926	3.155266	3.039631	0.115635
P8	1.91696	0.845672	1.071288	2.908992	1.113048	1.795944	2.854265	1.696854	1.15741
P9	2.85163	0.776438	2.075192	2.390113	0.250256	2.139857	2.903073	0.595902	2.307171
P10	1.086396	1.283209	-0.19681	2.067972	1.750523	0.317448	3.199787	3.222222	-0.02244
P11	2.355229	1.533753	0.821476	1.715378	2.070753	-0.35537	2.12439	2.209612	-0.08522
P12	3.394616	2.140621	1.253995	2.774139	3.286918	-0.51278	3.525314	2.861944	0.663369
P13	2.432972	1.716055	0.716917	3.313245	2.01617	1.297075	3.609015	2.740789	0.868226
P14	2.584308	0.79155	1.792759	2.101908	0.892238	1.20967	2.638981	1.508319	1.130662

Figure 18. An example of the rearranged file for the statistical analysis. Gaze event Fixation, condition Large Display

7.2 Analysis

To understand the influence of the display size on the cognitive load we will answer to our research sub-questions:

1. How do changes in Cognitive Load influence fixations in Large and Small displays?
2. How do changes in Cognitive Load influence saccades in Large and Small displays?
3. How do changes in Cognitive Load influence blink in Large and Small displays?
4. How do changes in Cognitive Load influence pupil dilation in Large and Small displays?
5. How do the number of correct answers and results of NASA TLX questionnaire correlate with the results gained with the help of eye-tracking technology?

First, we will investigate how the changes in the cognitive load influence each eye-tracking measure on the large and on the small displays. For this, we will observe the differences between the measures recorded OnRun period (where the participant is looking for the element, higher cognitive load) and measures recorded OffRun period (where the participant is in resting state, lower cognitive load). Later, we will investigate the differences between

the eye-tracking measures recorded on the large display and measures recorded on the small display. Finally, we will discuss how the display size affects the cognitive load.

Differences between two related groups, such as the differences between gaze measures recorded during OnRun and OffRun, and the differences between gaze measurements recorded for the large and small displays have been investigated using dependent T-Test.

The dependent T-Test is a statistical test and compares the means between two related groups. It is a parametric test which means that before conducting the test, one has to check the requirements for the data and make sure if the dependent t-test is a right choice.

Dependent T-test has four assumptions which the collected data has to fulfill [35]:

- **Assumption 1**: The dependent variable should be measured on a continuous scale. This means the data is measured either interval or ratio level. The gaze events are measured in interval level and fulfill the first assumption.
- **Assumption 2**: The independent variable should consist of the two related groups and the same subjects are presented in both groups. In the first phase of the analysis, we will study the differences of gaze measurements between OnRun and OffRun periods, which means that for both periods we collected the same gaze measurement (number of fixations, number of saccades, number of blinks and pupil dilation) from the same population. In the second phase of the analysis, we will study the differences of gaze measurements recorded on the large and on the small displays. We collect the same gaze measurements for both displays and from the same population. Our data fulfills the second assumption.
- **Assumption 3**: There should be no significant outlier in the differences between the two related groups. An outlier is a data point which is different from the rest of the data. This assumption is important since the outlier can affect the distribution of the data and lead to the invalid results. However, defining the significant outlier may be subjective and depend on the study.

  Outliers can be mild and extreme [36]. We interpreted the phrase in Assumption 3 - **significant outlier** to the extreme outlier, which means that our data should not

contain any extreme outlier. We used boxplot[11] to define the outlier in our data. First, we calculated the differences for the related groups as follows:

$$DiffRun_i = Run_iON - Run_iOFF \qquad i \in \{1,10\}$$

$$DiffRun_i = Run_iLargeDisplay - Run_iSmallDisplay \qquad i \in \{1,10\}$$

The mild outlier is the outlier which is located between the inner and outer fence of the boxplot on either side [37]. Inner and outer fences of the boxplot calculated as follows:

Lower inner fence: $Q1 - 1.5 \cdot IQ$

Upper inner fence: $Q3 + 1.5 \cdot IQ$

Lower outer fence: $Q1 - 3 \cdot IQ$

Upper outer fence: $Q3 + 3 \cdot IQ$

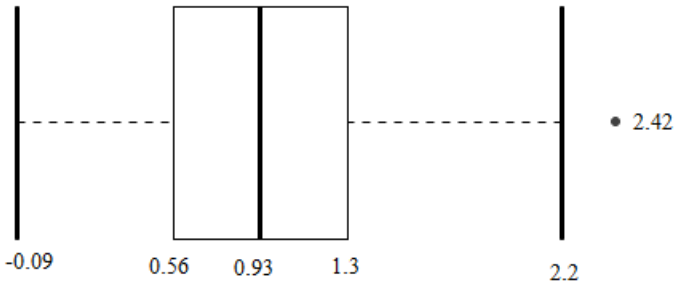


Figure 19. Boxplot of the differences of the related groups for Run 9. Metric fixation, Condition Small display. Mild outlier is the data collected from the participant id=33.

From the Figure 19, one can see that the data has an outlier. The outlier is investigated as follows:

[11] Boxplot is method for graphically describing the numerical data through their quartiles.

$$\text{Upper inner fence: } 1.3 + 1.5 \cdot (1.3 - 0.56) = 2.41$$

$$\text{Upper outer fence: } 1.3 + 3 \cdot (1.3 - 0.56) = 3.52$$

Data point 2.42 is a mild outlier since it is between the upper inner fence and upper outer fence.

It is also important to investigate the reason for the occurrence of an outlier. The most common reasons are missing factor during data collection, errors while data entry, incorrect data preparation process and naturally occurring outliers [38]. It is not uncommon that collected data from the eye-tracking experiment contains the outliers.

Our data contains outliers as well, however, they all are mild outliers. We collected the data in the same circumstances for all participants and conducted data processing process the same for all data with the help of programming language. Which means that all mild outliers in the data set are naturally occured. For this reason, we decided to keep them in the dataset.

- **Assumption 4**. The distribution of the differences in the dependent variable between two related groups should be approximately normally distributed. As in Assumption 3, we had to check if the *DiffRun* is normally distributed.
  As a normality test, we used Shapiro-Wilk test. The null-hypotheses for this test is that the data is normally distributed. If the p-value is lower than the chosen alpha, then the null-hypotheses is rejected, which means the data is not normally distributed. If the p-value is higher than the chosen alpha, then the null-hypotheses cannot be rejected, which means the data is normally distributed [39]. We have chosen the alpha as $\alpha = 0.05$.
  Normality can be tested with the graphical method such as Q-Q plot (Quantile-Quantile plot) as well. However, graphical methods are an informal way of assessing the normality.
  In Q-Q plot, if the points follow a linear pattern then the data is normally distributed and if the data points follow a non-linear pattern then the data is not normally distributed (see Figure 20).

In a real-world quantitative data analysis, it is common that in some cases the last assumption of the T-Test is violated. In this case, instead of T-Test, another statistical test has to be used which does not require the data to be normally distributed.

For the analysis, we used Wilcoxson signed-rank test which is used to compare means of two related groups as well. Before using this test, it is important to check if the data fulfills the assumptions of the test. Wilcoxson signed-rank test has three assumptions [35]. The first two assumptions are the same as the Assumption 1 and Assumption 2 of the dependent T-Test. Therefore, we will only discuss the last assumption of the Wilcoxson signed-rank test:

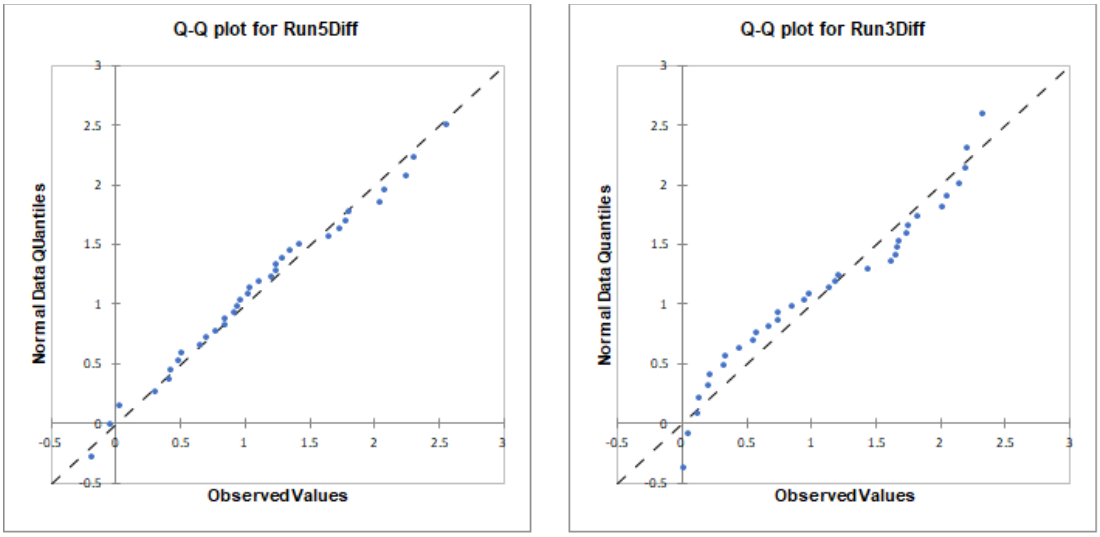


Figure 20. Q-Q Plot of the differences of the related groups in Run5 and the Run3, metric fixation, condition small display. Run5Diff is normally distributed (Shapiro-Wilk: p=0.821). Run3Diff is not normally distributed (Shapiro-Wilk: p=0.035).

- **Assumption 3**. The distribution of the differences between the two related groups needs to be symmetrical in shape. We used skewness to measure the degree of the symmetry. Distribution can be symmetric, skew to the right and skew to the left. If the skewness is near to zero, then the distribution is symmetric. If the skewness is smaller than zero then the distribution is skew to the left, otherwise to the right. It can be considered that the skewness higher than one in absolute value is highly skew,

skewness between 0.5 and 1 is moderately skew, and the skewness between 0 and 0.5 is fairly symmetric [40]. In our analysis, we considered the distribution with the absolute skewness between 0 and 0.5 as symmetrical in shape.

Symmetricity can be tested with a graphical method such as boxplot. However, these methods are an informal way of testing the symmetricity (see Figure 21).

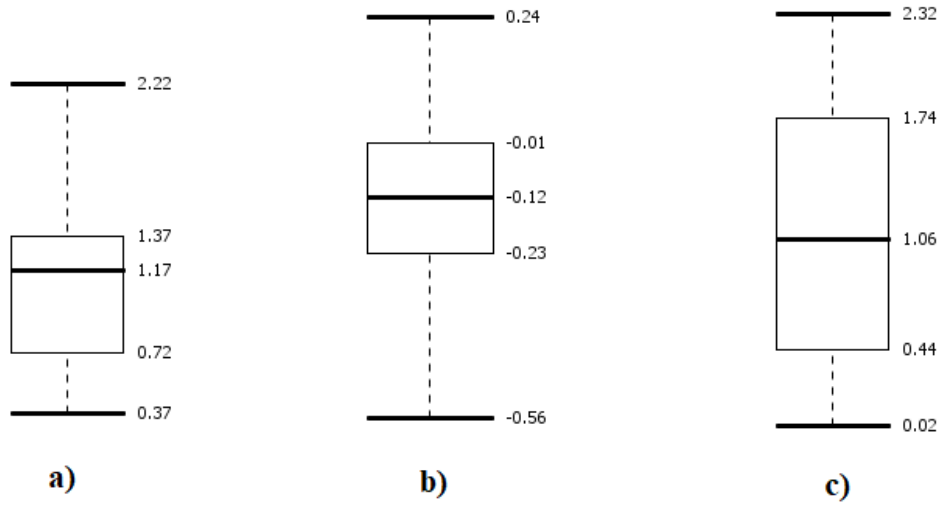


Figure 21 a) Boxplot visualization of the difference of the related groups in Run 3. Metric: fixation. Condition: small display. One can observe asymmetrical distribution. Skewness: -1.087 – highly skew. b) Boxplot visualization of the difference between the related groups in Run 5. Metric: blink. Condition: large display. With the boxplot visualization, it is difficult to observe if the distribution is fairly symmetric or skew. Skewness: -0.677 – moderately skew. c) Boxplot visualization of the differences of the related groups in Run 2. Metric: fixation. Condition: small display. One can observe symmetrical distribution. Skewness: -0.092 – fairly symmetric.

In the real-world data, it is not unlikely that the assumption about the symmetricity is violated. In this case, differences between related groups can be calculated with the help of non-parametric test such as Sign test.

Sign test has four assumptions [35]. First two assumptions are the same as the Assumption 1 and Assumption 2 of the T-test, therefore, we will discuss Assumption 3 and Assumption 4 of the Sign test:

- **Assumption 3**: The paired observations for each participant need to be independent, that is, the value of one participant must not influence the value of the other participant. In our experiment, each participant had own time slot for participation

and the data from each participant prepared separately. The values from one participant did not influence the values of the other participants. The data fulfills this assumption.

- **Assumption 4**. The differences between two related groups are from a continuous distribution. The *DiffRun* can take any value and it is continuous. The data fulfills this assumption.

In the following section, we will present the result of the analysis. Although we conduct the experiment with 35 participants, we analyzed the data from 34 participants. The data collected from the participant id=4 contained a large amount of missing values and the eye-tracker did not collect the size of pupil throughout the experiment. In this regard, we excluded the data from this participant from the analysis.

8. Results

In this section, we will present the results gained with the help of eye-tracking technology, the results gained through the NASA TLX questionnaire and the performances of the participants. We assume that by increasing the number of distractors after each run the cognitive load increases. Our goal is to understand how does the number of gaze events change when the cognitive load increases. We will investigate the differences between the gaze events during the active search (OnRun period – higher cognitive load) and resting state (OffRun – lower cognitive load). After this, we will investigate the difference between gaze events collected on the large display and small display.

Additionally, we will discuss how the objective results (results obtained with the eye-tracking technology), the performance of the participant and subjective results (results obtained through the NASA TLX questionnaire) match.

Results for each measurement, such as fixation, saccade, blink and pupil dilation are reported separately.

8.1 Fixation

Many researchers have been investigating the relation between fixation and cognitive state [20] [10] and they showed higher load in working memory leads to an increment in the number of fixation. And we want to know if the fixation can help us to detect the influence of the display size on the cognitive load.

In this section, we will answer to our 1st research sub-question: How do changes in Cognitive Load influence fixation in Large and Small displays? In order to answer to this question, we will study how the number of fixations per second changes from OnRun period, where participant is looking for the target element (higher cognitive load) to OffRun period, where participant has already found the target element (or the lap duration is over) and waiting for the next lap to start. We will present results for each condition separately.

Large Display

We used dependent T-Test to compare OnRun and OffRun periods. Before conducting this test, one has to check if the differences between the two related groups are normally distributed. We tested normality using Shapiro-Wilk test. Table 2 presents the results of dependent T-Test in run level. For each run, we ran a T-Test for the average values of fixations per second. These values are averaged by participants recorded during OnRun and OffRun periods. In our tables, we used colours according to the relevance. Green colour means that the data does not violate the assumption of the corresponding statistical test. Blue colour means that the data is statistically significant, otherwise, the result of the statistical test has a red background. We will follow this colouring schema in all tables where we conduct the statistical test.

Additionally, tables present descriptive statistics, such as a mean number of gaze events and standard deviation.

Data is normally distributed if the *p*-value of Shapiro-Wilk test is $p \geq 0.05$ (green background) and not normally distributed if *p*-value is $p < 0.05$. Data is statistically significant if the *p*-value of dependent T-Test (as well as Wilcoxson signed-rank test and Sign test) is $p \leq 0.5$ (blue background) and not statistically significant if the *p*-value is $p > 0.05$ (pink background).

Observing Table 2, we can see that *p*-value of Shapiro-Wilk test is $p \geq 0.05$ for every run, which means that the data is normally distributed. This allows us to conduct dependent T-Test. The results of dependent T-Test are statistically significant for every run, which means that the gaze event fixation differs significantly between OnRun and OffRun periods.

Additionally, a positive number of *t*-values indicate a higher number of fixations per second in the OnRun period. This is the result of the higher cognitive load where participants have to focus on more items and move eyes more repeatedly during the active search.

Table 2. Results of the dependent T-Test for the gaze event Fixation, condition Large Display. Normality is checked with the Shapiro-Wilk test.

Nº Fixation/s -Large display		Descriptive Statistic		Shapiro-Wilk Test	Dependent T-Test	
		Mean Nº Fix/s	Std. Deviation	p-value	t-value	p-value
Pair 1	**R1on**	2.387	0.510	0.305	8.697	**0.000**
	R1off	1.241	0.704			
Pair 2	**R2on**	2.696	0.479	0.539	9.337	**0.000**
	R2off	1.402	0.712			
Pair 3	**R3on**	2.843	0.477	0.878	10.535	**0.000**
	R3off	1.715	0.831			
Pair 4	**R4on**	2.985	0.405	0.790	12.282	**0.000**
	R4off	1.662	0.715			
Pair 5	**R5on**	2.954	0.344	0.554	9.772	**0.000**
	R5off	1.805	0.811			
Pair 6	**R6on**	2.941	0.431	0.220	7.544	**0.000**
	R6off	1.886	0.764			
Pair 7	**R7on**	2.913	0.436	0.805	10.986	**0.000**
	R7off	1.746	0.736			
Pair 8	**R8on**	2.860	0.385	0.885	9.006	**0.000**
	R8off	1.770	0.677			
Pair 9	**R9on**	2.820	0.426	0.475	8.324	**0.000**
	R9off	1.752	0.754			
Pair 10	**R10on**	2.757	0.417	0.368	8.408	**0.000**
	R10off	1.916	0.608			

Studying Table 2, we can conclude that the number of fixations is higher in OnRun periods ($t > 0$) and these differences are statistically significant for all runs. Based on these results, we can say that the fixation is an indicator of the level of cognition in condition Large display.

Small Display

For condition small display we have followed the same procedure as for condition large display. Normality is checked with Shapiro-Wilk test. If the data is normally distributed, we conducted dependent T-Test to compare two related groups. If the data is not normally distributed, we conducted an alternative statistical test which does not require the data to be normally distributed. For this case, we used Wilcoxson signed-rank test. This test requires the data to be symmetrical in shape. We checked the symmetricity by calculating the skewness of the data, as we already discussed in the previous chapter. If the skewness is $|skewness| \leq 0.5$, we can conduct Wilcoxson signed-rank test. However, if the data violets this requirement, we conducted non-parametrical Sign test to compare two related groups.

From Table 3, we can observe that except Run 2 and Run 3 the data is normally distributed, which allows us to conduct dependent T-Test. For Run 2 we conducted Wilcoxson test, as the skewness of the data is $|skewness| \leq 0.5$ (fairly symmetrical). For Run 3, we conducted Sign test, since the data violated the assumptions of the dependent T-Test and Wilcoxson signed-rank test.

Table 3 shows that the number of fixations per second is higher in OnRun period than the OffRun period ($t > 0,\ z > 0$) and this difference is statistically significant for all runs.

From condition Large display and Small display, we have learned that the difference between the data collected during the OnRun period (higher cognitive load) and OffRun (lower cognitive load) period is statistically significant. Positive *t*-values and *z*-score show that the number of fixations per second is higher on the OnRun period. Using this knowledge in the next section, we will conduct the statistical analysis to understand how the display size influence the level of cognition.

Table 3. Results of the dependent T-Test, Wilcoxson test and Sign test for the gaze event Fixation, condition Small Display. Normality is checked with the Shapiro-Wilk test.

Nº Fixation/s -Small display		Descriptive Statistic		Shapiro-Wilk Test	Dependent T-Test		Sym-met-ricity	Wilcoxson Test		Sign Test
		Mean Nº Fix/s	Std. Deviation	p-value	t-value	p-value	Skewness	z-score	p-value	p-value
Pair 1	**R1on**	2.519	0.474	0.497	9.540	**0.000**				
	R1off	1.375	0.719							
Pair 2	**R2on**	2.491	0.452	0.021			-0.092	4.873	**0.000**	
	R2off	1.424	0.644							
Pair 3	**R3on**	2.805	0.469	0.035			-1.087			**0.000**
	R3off	1.693	0.682							
Pair 4	**R4on**	2.708	0.421	0.396	9.394	**0.000**				
	R4off	1.699	0.715							
Pair 5	**R5on**	2.912	0.433	0.821	9.386	**0.000**				
	R5off	1.803	0.805							
Pair 6	**R6on**	2.744	0.407	0.221	10.322	**0.000**				
	R6off	1.659	0.592							
Pair 7	**R7on**	2.829	0.422	0.113	9.240	**0.000**				
	R7off	1.858	0.697							
Pair 8	**R8on**	2.854	0.452	0.683	9.621	**0.000**				
	R8off	1.803	0.739							
Pair 9	**R9on**	2.836	0.317	0.504	9.244	**0.000**				
	R9off	1.858	0.676							
Pair 10	**R10on**	2.821	0.330	0.134	8.881	**0.000**				
	R10off	1.943	0.565							

Large vs. Small display

As we already discussed, fixation can be used as an indicator of the level of cognition in both conditions. And we concluded that higher cognitive load leads to a higher number of fixations per second. Based on these findings, we can conduct another test to understand if there is a difference in cognitive load while participant performing the task on the large and small displays.

In this regard, we will compare the mean values of the fixations per second between the large and small displays. We will do this comparison for the data collected only in the OnRun period. Since we are interested in the differences in the higher cognitive load.

Table 4 shows the results of the comparison between the large and small displays. Except Run 9 (Shapiro-Wilk test: $p = 0.035$) all data is normally distributed, and we are allowed to conduct the dependent T-Test. For the Run 9, we conducted Sign test, since the data is moderately skew ($skewness = -0.876$).

Except Run 4 and Run 6, the differences between the number of fixations per second are not statistically significant. Which means, not every participant had a higher number of fixations on the large display than the small display or vice versa.

Additionally, from the *t*-values of the dependent T-Test one can observe that only for Run 1 and Run 10 the value is negative. This means that the number of fixations per second is higher on the small display for Run 1 and Run 10. For the rest of the runs, performing the task on the large display produced more fixation than the small display.

Although, the number of fixations is higher on the large display except the Run 1, Run 9 (regarding the mean values) and Run 10, these differences are statistically significant only for Run 4 and Run 6. Therefore, we cannot conclude that performing the task on the large display yields more number of fixations than on the small display for each difficulty level of the task (for each run). Thus, the display size has no significant influence on the cognitive load at the run level for the measurement fixation.

Table 4. Results of the dependent T-Test and Sign test for the gaze event Fixation, Large vs. Small display comparison. Normality is checked with the Shapiro-Wilk test.

Nº Fixation/s -Large vs. Small display		Descriptive Statistic		Shapiro-Wilk Test	Dependent T-Test		Sym-met-ricity	Wilcoxson Test		Sign Test
		Mean Nº Fix/s	Std. Deviation	p-value	t-value	p-value	Skewness	z-score	p-value	p-value
Pair 1	**R1L**	2.387	0.510	0.698	-1.160	**0.254**				
	R1S	2.519	0.474							
Pair 2	**R2L**	2.696	0.479	0.299	1.974	**0.057**				
	R2S	2.491	0.452							
Pair 3	**R3L**	2.843	0.477	0.457	0.385	**0.703**				
	R3S	2.805	0.469							
Pair 4	**R4L**	2.985	0.405	0.741	4.037	**0.000**				
	R4S	2.708	0.421							
Pair 5	**R5L**	2.954	0.344	0.603	0.475	**0.638**				
	R5S	2.912	0.433							
Pair 6	**R6L**	2.941	0.431	0.879	2.061	**0.047**				
	R6S	2.744	0.406							
Pair 7	**R7L**	2.913	0.436	0.770	1.016	**0.317**				
	R7S	2.829	0.422							
Pair 8	**R8L**	2.860	0.385	0.496	0.058	**0.954**				
	R8S	2.854	0.452							
Pair 9	**R9L**	2.820	0.426	0.035			-0.876			**0.392**
	R9S	2.836	0.317							
Pair 10	**R10L**	2.758	0.417	0.291	-0.899	**0.376**				
	R10S	2.821	0.330							

However, it is interesting to study the difference of the number of fixations between the large and small display not for every run, but for the whole task. Table 5 shows the comparison of the number of fixations per second between the large and small displays.

Table 5. Results of the dependent T-Test for the gaze event Fixation averaged for the whole task. Large vs. Small display comparison. Normality is checked with the Shapiro-Wilk test.

Σ º Fixation/s -Large vs. Small display	Descriptive Statistic		Shapiro-Wilk Test	Dependent T-Test	
	Mean Nº Fix/s	Std. Deviation	p-value	t-value	p-value
Large Display	2.816	0.248	0.051	1.659	**0.107**
Small Display	2.754	0.237			

From Table 5, we can observe that the data is normally distributed, and we can conduct the dependent T-Test. The *t*-value is positive which means the mean number of fixations per second in condition Large display is higher than in condition Small display. However, the difference is not statistically significant. Therefore, using the measure fixation, we cannot conclude that performing the task on the Large display yields higher cognitive load.

8.2 Saccade

There have been many investigations to study relation between the cognitive load and saccade and the result of the researches show that higher cognitive load leads to an increment in the number of saccades [41][10]. However, we want to know if the saccade can be used as an indicator to measure cognitive load in different sized display and how this eye movement can help us to understand the influence of the display size on the cognitive load.

In this section, we will answer to our 2nd research sub-question: How do changes in Cognitive Load influence saccade in Large and Small displays? In order to answer to this question, first we will study how the number of saccades per second changes from OnRun period, where participant is looking for the target element (higher cognitive load) to OffRun period, where participant has already found the target element (or the lap duration is over) and waiting for

the next lap to start (lower cognitive load). We will present results for each condition separately.

Large display

For condition Large display, we conducted dependent T-Test to understand how the number of fixations differs between OnRun and OffRun periods. Before conducting the statistical test, we checked the normality of the data with the Shapiro-Wilk test.

Table 6 shows the result of the comparison between OnRun and OffRun periods. For every run the data is normally distributed which allows us to conduct the dependent T-Test. The difference between the number of saccades per second of the OnRun and OffRun period is statistically significant for every run.

Additionally, the number of positive *t*-values show that the number of fixations is higher during the OnRun period than an OffRun period. This is the result of the higher cognitive load where participants need to move eyes more intensively while active search.

Studying Table 6, we can conclude that the number of saccades is higher in the OnRun period ($t > 0$) and this difference is statistically significant for all runs. Based on these results, we can say that the saccade is an indicator of the level of cognition in condition Large display.

Table 6. Results of the dependent T-Test for the gaze event Saccade, condition Large Display. Normality is checked with the Shapiro-Wilk test.

Nº Saccade/s -Large display		Descriptive Statistic		Shapiro-Wilk Test	Dependent T-Test	
		Mean Nº Sacc/s	Std. Deviation	p-value	t-value	p-value
Pair 1	**R1on**	2.527	0.506	0.124	13.344	**0.000**
	R1off	1.051	0.635			
Pair 2	**R2on**	2.747	0.539	0.313	11.713	**0.000**
	R2off	1.234	0.682			
Pair 3	**R3on**	2.845	0.495	0.940	11.737	**0.000**
	R3off	1.561	0.822			
Pair 4	**R4on**	2.989	0.425	0.952	14.826	**0.000**
	R4off	1.541	0.695			
Pair 5	**R5on**	2.970	0.351	0.933	10.372	**0.000**
	R5off	1.753	0.816			
Pair 6	**R6on**	2.909	0.500	0.096	8.718	**0.000**
	R6off	1.734	0.724			
Pair 7	**R7on**	2.918	0.470	0.247	11.932	**0.000**
	R7off	1.670	0.767			
Pair 8	**R8on**	2.837	0.411	0.612	8.970	**0.000**
	R8off	1.690	0.717			
Pair 9	**R9on**	2.810	0.473	0.326	8.872	**0.000**
	R9off	1.692	0.777			
Pair 10	**R10on**	2.725	0.486	0.351	8.569	**0.000**
	R10off	1.786	0.663			

Small display

For condition Small display, we followed the same procedure as for condition Large display. Normality is checked with Shapiro-Wilk test. If the data is normally distributed, we conducted dependent T-Test to compare two related groups.

Table 7. Results of the dependent T-Test for the gaze event Saccade, condition Small Display. Normality is checked with the Shapiro-Wilk test.

Nº Saccade/s -Small display		Descriptive Statistic		Shapiro-Wilk Test	Dependent T-Test	
		Mean Nº Sacc/s	Std. Deviation	p-value	t-value	p-value
Pair 1	**R1on**	2.614	0.504	0.690	12.771	**0.000**
	R1off	1.149	0.649			
Pair 2	**R2on**	2.573	0.483	0.543	14.262	**0.000**
	R2off	1.198	0.637			
Pair 3	**R3on**	2.804	0.514	0.096	10.480	**0.000**
	R3off	1.532	0.652			
Pair 4	**R4on**	2.691	0.419	0.162	10.781	**0.000**
	R4off	1.533	0.689			
Pair 5	**R5on**	2.902	0.487	0.062	10.690	**0.000**
	R5off	1.651	0.805			
Pair 6	**R6on**	2.736	0.459	0.572	13.119	**0.000**
	R6off	1.505	0.534			
Pair 7	**R7on**	2.781	0.484	0.673	9.083	**0.000**
	R7off	1.778	0.718			
Pair 8	**R8on**	2.819	0.509	0.869	9.207	**0.000**
	R8off	1.754	0.772			
Pair 9	**R9on**	2.799	0.363	0.929	8.867	**0.000**
	R9off	1.817	0.662			
Pair 10	**R10on**	2.781	0.400	0.291	8.500	**0.000**
	R10off	1.802	0.616			

From Table 7, we can observe that the data is normally distributed for every run, which allows us to conduct the dependent T-Test. The results of the dependent T-Test are statistically significant for every run. This means that the number of saccades is significantly different between OnRun and OffRun periods.

Furthermore, positive *t*-values show that the number of saccades is higher in the OnRun period than the OffRun period. This is the result of the need to move eyes more intensively while active search (higher cognitive load).

Studying Table 7, we can conclude that the number of saccades is higher in OnRun periods ($t > 0$) and these differences are statistically significant for all runs. Based on these results, we can say that the number of saccades is an indicator of the level of cognition in condition Small display.

From the results of the condition Large display and Small display, we have learned that the differences between the data collected during the OnRun period (higher cognitive load) and OffRun period (lower cognitive load) are statistically significant. Positive *t*-values show that the number of saccades per second is higher on the OnRun period. Using this knowledge in the next section, we will conduct the statistical analysis to understand how the number of saccades per second change between the large and small displays and if the display size has any effect on the cognitive load.

Large vs. Small display

As we discussed in previous sections, saccade can be used as an indicator of the level of cognition in both conditions. And we concluded that higher cognitive load leads to a higher number of saccades. Based on these findings, we can conduct another test to understand if there is a difference in cognitive load while participant performing the task on the large and small displays.

In this regard, we will compare the mean values of the saccades per second between the large and small displays. For comparison, we will use the data collected only in the OnRun period.

Table 8 shows the results of the comparison between the large and small displays. Except Run 6, the data is normally distributed, and we can conduct the dependent T-Test. For the Run 6, we conducted Sign test, since the data violates the assumption of symmetricity of the Wilcoxson signed-rank test.

Table 8. Results of the dependent T-Test and Sign test for the gaze event Saccade, Large vs. Small display comparison. Normality is checked with the Shapiro-Wilk test.

Nº Saccades/s -Large vs. Small Display		Descriptive Statistic		Shapiro-Wilk Test	Dependent T-Test		Sym-met-ricity	Wilcoxson Test		Sign Test
		Mean Nº Sacc/s	Std. Deviation	p-value	t-value	p-value	Skew ness	z-score	p-value	p-value
Pair 1	**R1L**	2.527	0.506	0.831	-0.824	**0.416**				
	R1S	2.614	0.504							
Pair 2	**R2L**	2.748	0.539	0.217	1.618	**0.115**				
	R2S	2.573	0.483							
Pair 3	**R3L**	2.845	0.495	0.358	0.402	**0.690**				
	R3S	2.804	0.514							
Pair 4	**R4L**	2.989	0.425	0.640	4.543	**0.000**				
	R4S	2.691	0.419							
Pair 5	**R5L**	2.970	0.351	0.058	0.824	**0.416**				
	R5S	2.902	0.487							
Pair 6	**R6L**	2.909	0.500	0.039			0.844			**0.121**
	R6S	2.736	0.459							
Pair 7	**R7L**	2.918	0.470	0.636	1.599	**0.119**				
	R7S	2.781	0.484							
Pair 8	**R8L**	2.837	0.411	0.050	0.194	**0.484**				
	R8S	2.819	0.509							
Pair 9	**R9L**	2.810	0.473	0.160	0.163	**0.872**				
	R9S	2.799	0.363							
Pair 10	**R10L**	2.725	0.486	0.239	-0.818	**0.419**				
	R10S	2.781	0.400							

Except Run 4, the differences between the number of saccades per second are not statistically significant. Additionally, from the *t*-values of the dependent T-Test one can observe that only for Run 1 and Run 10 the value is negative. This means that the number of saccades per second is higher on the small display for Run 1 and Run 10. For the rest of the runs, performing the task on the large display produced more saccades than the small display.

Although, the number of fixations is higher on the large display except the Run 1, Run 6 (regarding the mean values) and Run 10, these differences are statistically significant only for Run 4. Therefore, we cannot conclude that the size of the display has any influence on the cognitive load at each difficulty level of the task.

It is interesting to study the difference of the mean number of saccades between the large and small display not for every run, but for the whole task. Table 9 shows the comparison of the number of saccades per second between the large and small displays.

Table 9. Results of the Sign test for the gaze event Saccade averaged for the whole task. Large vs. Small display comparison.

Σ º Saccade/s Large vs. Small displays	**Descriptive Statistic**		**Shapiro-Wilk Test**	**Dependent T-Test**		**Symmet -ricity**	**Wilcoxson Test**		**Sign Test**
	Mean Nº Saccade/s	Std. Deviation	p-value	t-value	p-value	Skewness	z-score	p-value	p-value
Large Display	2.828	0.407	0.003			1.451			**0.000**
Small Display	2.750	0.322							

From Table 9, we can observe that the data is not normally distributed ($p = 0.003$) and highly skew ($skewness = 1.451$), and we cannot conduct the dependent T-Test and Wilcoxson signed-rank test. Therefore, we conducted non-parametric Sign test. The result of the Sign test is statistically significant, which means that the number of saccades between the large and small displays differs significantly.

Additionally, regarding the mean values from the Table 9, we can conclude that the number of saccades produced on the large display is significantly higher than on the small display.

Using gaze event fixation, we could not find the statistically significant difference between our two conditions on a different level of cognition (for every run of the task). However, the mean number of the saccade for the whole task differs significantly between the large and small displays. Thus, performing the task on the large display produces more saccades than

on the small display. Respectively, using the measure saccade, we can conclude that large display yields higher cognitive load at the task level.

8.3 Blink

Many researches show that the number of blinks has a direct relation to the cognitive load and the higher cognitive load leads to the lower number of blinks [20][41]. We want to investigate how the number of blinks change while increasing the cognitive load on both large and small displays and can we use the blink rate to detect the influence of the display size on the cognitive load.

In this section, we will answer to our 3rd research sub-question: How do changes in Cognitive Load influence blink in Large and Small displays? To answer to this question, we will study how the number of blinks per second changes from OnRun period, where participant is looking for the target element (higher cognitive load) to OffRun period, where participant has already found the target element (or the lap duration is over) and waiting for the next lap to start (lower cognitive load). We will present results for each condition separately.

Large display

For condition Large display, we conducted statistical tests to compare two related groups. Normality is checked with the Shapiro-Wilk test. If the data is normally distributed, we conducted dependent T-Test. If the data is not normally distributed but fairly symmetric, we conducted Wilcoxson signed-rank test. If the data violates conditions of both parametric tests, non-parametric Sign test is used to compare two related groups.

Table 10 shows the results of the comparison between OnRun and OffRun periods for the condition Large display. The data is not normally distributed except in Run 8, Run 9 and Run 10. For the last three runs, we conducted dependent T-Test. The data in Run 3 is fairly symmetrical in shape and we are allowed to conduct Wilcoxson signed-rank test. For the rest of the runs, we conducted the Sign test.

Table 10. Results of the dependent T-Test, Wilcoxson signed-rank test and the Sign test for the gaze event Blink, condition Large Display. Normality is checked with the Shapiro-Wilk test.

Nº Blink/s -Large display		Descriptive Statistic		Shapiro-Wilk Test	Dependent T-Test		Symmet-ricity	Wilcoxson Test		Sign Test
		Mean Nº Blink/s	Std. Deviation	p-value	t-value	p-value	Skewness	z-score	p-value	p-value
Pair 1	**R1on**	0.063	0.142	0.001			-1.404			**0.000**
	R1off	0.242	0.219							
Pair 2	**R2on**	0.085	0.155	0.004			-1.218			**0.000**
	R2off	0.233	0.188							
Pair 3	**R3on**	0.090	0.157	0.033			0.111	-3.867	**0.000**	
	R3off	0.216	0.173							
Pair 4	**R4on**	0.090	0.143	0.008			-1.384			**0.000**
	R4off	0.276	0.224							
Pair 5	**R5on**	0.089	0.149	0.034			-0.677			**0.000**
	R5off	0.232	0.211							
Pair 6	**R6on**	0.108	0.171	0.006			-1.064			**0.000**
	R6off	0.288	0.241							
Pair 7	**R7on**	0.089	0.150	0.012			-1.204			**0.000**
	R7off	0.255	0.236							
Pair 8	**R8on**	0.095	0.146	0.120	-4.509	**0.000**				
	R8off	0.272	0.271							
Pair 9	**R9on**	0.109	0.169	0.948	-4.933	**0.000**				
	R9off	0.237	0.195							
Pair 10	**R10on**	0.096	0.131	0.415	-6.116	**0.000**				
	R10off	0.284	0.194							

The results of the tests are statistically significant for all 10 runs. This means that the data collected OnRun and OffRun periods differ significantly.

Additionally, negative *t*-values and *z*-scores, and the differences in the mean values show that the number of blinks is higher in the OffRun period than the OnRun period. This is the result of the need to keep eyes open as much as possible during the active search to find the target item and blink to moisten the eyes during the resting the resting state.

Studying Table 10, we can conclude that the number of blinks during the active search (OnRun) is lower than during the resting state (OffRun) and this difference is statistically significant for all runs. Based on these results, we can say that the number of blinks is an indicator of the level of cognition in condition Large display.

Small display

For condition Small display, we followed the same procedure as for condition Large display. We conducted dependent T-Test and Sign test to compare two related groups. Normality is checked with Shapiro-Wilk test. The data collected for the Run 1, Run 3, Run 4, Run 7, Run 8 are normally distributed, accordingly, we can conduct dependent T-Test. The data collected for the rest of the runs violates the assumption of the symmetricity of the Wilcoxson signed-rank test, therefore we conducted Sign test to compare OnRun and OffRun periods.

Table 11 shows the result of the tests for condition Small display. Except the Run 9, the results are statistically significant. This means that the number of blinks collected OnRun and OffRun periods differ significantly. Additionally, negative *t*-values of the dependent T-Test and mean values from the descriptive statistics show that the number of blinks is higher in OffRun period than in OnRun period and these differences are statistically significant for all runs except Run 9. Although, from the mean values of the Run 9, we can see that number of blinks in OffRun period is higher, however, this was not enough for the threshold alpha of the Sign test ($\alpha = 0.05$, $p = 0.071$), accordingly, the difference in Run 9 is not statistically significant.

Studying Table 11, we can conclude that the number of blinks during the active search (OnRun) is lower than during the resting state (OffRun) and this difference is statistically

significant for all runs. Based on these results, we can say that the number of blinks is an indicator of the level of cognition in condition small display.

Table 11. Results of the dependent T-Test and Sign test for the gaze event Blink, condition Small display. Normality is checked with the Shapiro-Wilk test.

Nº Blink/s -Small display		Descriptive Statistic		Shapiro-Wilk Test	Dependent T-Test		Symmet-ricity	Wilcoxson Test		Sign Test
		Mean Nº Blink/s	Std. Deviation	p-value	t-value	p-value	Skewness	z-score	p-value	p-value
Pair 1	**R1on**	0.080	0.127	0.325	-5.490	**0.000**				
	R1off	0.292	0.245							
Pair 2	**R2on**	0.095	0.167	0.013			-1.284			**0.000**
	R2off	0.302	0.236							
Pair 3	**R3on**	0.112	0.148	0.842	-5.575	**0.000**				
	R3off	0.281	0.222							
Pair 4	**R4on**	0.108	0.172	0.060	-4.498	**0.000**				
	R4off	0.249	0.178							
Pair 5	**R5on**	0.107	0.150	0.013			-1.198			**0.000**
	R5off	0.259	0.238							
Pair 6	**R6on**	0.103	0.141	0.014			-1.266			**0.000**
	R6off	0.277	0.254							
Pair 7	**R7on**	0.134	0.185	0.421	-3.502	**0.001**				
	R7off	0.251	0.196							
Pair 8	**R8on**	0.106	0.123	0.065	-4.625	**0.000**				
	R8off	0.249	0.206							
Pair 9	**R9on**	0.093	0.110	0.001			-1.702			**0.071**
	R9off	0.206	0.215							
Pair 10	**R10on**	0.100	0.139	0.039			-0.907			**0.000**
	R10off	0.282	0.236							

From the results of the condition Large display and Small display, we have learned that the differences between the data collected during the OnRun period (higher cognitive load) and OffRun period (lower cognitive load) are statistically significant and the number of blinks per second is higher on the OffRun period. Using this knowledge in the next section, we will conduct the statistical analysis to understand how the number of blinks changes between the large and small displays.

Large vs. Small display

As we already discussed, the number of blinks can be used as an indicator of the level of cognition. And we concluded that higher cognitive load leads to a lower number of blinks. Using this knowledge, we can conduct another test to understand if there is a difference in cognitive load while participant performing the task on the large and small displays.

In this regard, we will compare the mean values of the blinks per second between the large and small displays. For comparison, we will use the data collected only in the OnRun period.

Table 12 shows the results of the comparison of the number of blinks per second collected during the OnRun period between the large and small displays. Except Run 3 and Run 6, the data is not normally distributed, accordingly, we conducted dependent T-Test for Run 3 and Run 6. The data collected in Run 10 is symmetrical in shape ($skewness = 0.296$) which allows us to conduct Wilcoxson signed-rank test. For the rest of the runs, we conducted non-parametric Sign test.

Except for Run7, the results of the tests are not statistically significant. Positive *t*-value of the dependent T-Test for Run 6 and the mean values of the descriptive statistics for Run 9 show that performing the task on the large display produced more blinks than the small display. For the rest of the runs, the number of blinks is higher in the condition small display. However, this difference is statistically significant only for the Run 7.

Studying Table 12, we cannot conclude that the display size has any influence on the cognitive load at the run level.

Table 12. Results of the dependent T-Test, Wilcoxson signed-rank test and Sign test for the gaze event Blink, Large vs. Small display comparison. Normality is checked with the Shapiro-Wilk test.

Nº Blink/s -Large vs. Small Display		Descriptive Statistic		Shapiro-Wilk Test	Dependent T-Test		Symmetricity	Wilcoxson Test		Sign Test
		Mean Nº Blink/s	Std. Deviation	p-value	t-value	p-value	Skewness	z-score	p-value	p-value
Pair 1	**R1L**	0.064	0.142	0.000			1.034			**0.238**
	R1S	0.080	0.127							
Pair 2	**R2L**	0.085	0.155	0.000			-0.974			**0.824**
	R2S	0.096	0.167							
Pair 3	**R3L**	0.090	0.157	0.069	-0.961	**0.344**				
	R3S	0.112	0.148							
Pair 4	**R4L**	0.090	0.143	0.000			1.442			**0.122**
	R4S	0.108	0.172							
Pair 5	**R5L**	0.089	0.149	0.001			1.303			**0.150**
	R5S	0.107	0.150							
Pair 6	**R6L**	0.108	0.171	0.100	0.279	**0.782**				
	R6S	0.103	0.141							
Pair 7	**R7L**	0.089	0.150	0.002			-1.430			**0.005**
	R7S	0.134	0.185							
Pair 8	**R8L**	0.095	0.146	0.002			0.936			**0.122**
	R8S	0.106	0.123							
Pair 9	**R9L**	0.109	0.169	0.000			1.890			**1.000**
	R9S	0.094	0.110							
Pair 10	**R10L**	0.096	0.131	0.027			0.296	-0.167	**0.868**	
	R10S	0.100	0.139							

However, it is interesting to study the difference between the mean number of blinks of the large and small display not for every run, but for the whole task. Table 13 shows the

comparison of the number of saccades per second collected for the whole task between the large and small displays.

From Table 13, we can observe that the data is not normally distributed ($p = 0.023$) and we cannot conduct dependent T-Test. The data is fairly symmetric ($skewness = 0.418$) and we are allowed to conduct Wilcoxson signed-rank test.

Table 13. Results of the Wilcoxson signed-rank test for the gaze event Blink averaged for the whole task. Large vs. Small display comparison.

Σ º Blink/s Large vs. Small Display	**Descriptive Statistic**		**Shapiro-Wilk Test**	**Dependent T-Test**		**Symmet-ricity**	**Wilcoxson Test**	
	Mean Nº Blink/s	Std. Deviation	p-value	t-value	p-value	Skewness	z-score	p-value
Large Display	0.091	0.139	0.023			0.418	-2.126	**0.033**
Small Display	0.104	0.128						

The result of the Wilcoxson signed-rank test is statistically significant, which means that the number of blinks for the whole task differs significantly between the large and small displays. The negative *z*-score of the Wilcoxson signed-rank test shows that the number of blinks produced on the small display is higher than the number of blinks produced on the large display and this difference is statistically significant.

Using gaze event blink, we could not find the statistically significant difference between our two conditions at the different level of cognition (for every run of the task). However, the mean number of the blinks for the whole task differs significantly between the large and small displays. Respectively, using the measure blinks, we can conclude that large display yields higher cognitive load at the task level.

8.4 Pupil dilation

Pupil dilation is the most investigated eye movement and it has been shown that higher cognitive load leads to an increment of the pupil diameter [20][18][10]. But we want to know how the pupil dilation change while increasing the cognitive load in visual search and if these changes follow the same pattern for the large and small displays.

In this section, we will answer to our 4th research sub-question: How do changes in Cognitive Load influence pupil dilation in Large and Small displays? To answer to this question, we will investigate how the pupil dilation changes from the OnRun period, where the participant is in active search (higher cognitive load) to OffRun period, where the participant is in resting state (lower cognitive load). We will present results for each condition separately.

Large display

For the condition Large display, we conducted statistical tests to compare the pupil dilation recorded in OnRun and OffRun periods. If the data is normally distributed, we conducted dependent T-Test. If the data is not normally distributed but fairly symmetric, we conducted Wilcoxson signed-rank test. If the data violates the assumptions of both statistical tests, we conducted Sign test.

Table 14 shows the results of the comparison for the large display. The data collected in the Run 2, Run 3, Run 6, Run 7, Run 8 and Run 10 are normally distributed and we can conduct dependent T-Test. The data from the Run 1 and Run 9 are fairly symmetric which allows us to conduct Wilcoxson signed-rank test. For the rest of the runs, we conducted non-parametric Sign test.

The results of the tests are statistically significant for the Run 3, Run 6, Run 7 and Run 8. For the rest of the runs, the difference of the pupil dilation between our two conditions is not statistically significant.

Table 14. Results of the dependent T-Test, Wilcoxson signed rank-test and Sign test for the gaze event Pupil dilation, condition Large display. Normality is checked with the Shapiro-Wilk test.

Pupil dilation -Large display		Descriptive Statistic		Shapiro -Wilk Test	Dependent T-Test		Symmet-ricity	Wilcoxson Test		Sign Test
		Mean diam. pupil	Std. Deviation	p-value	t-value	p-value	Skewness	z-score	p-value	p-value
Pair 1	**R1on**	2.593	0.361	0.028			-0.086	1.556	**0.120**	
	R1off	2.571	0.371							
Pair 2	**R2on**	2.483	0.348	0.245	-0.421	**0.676**				
	R2off	2.492	0.366							
Pair 3	**R3on**	2.484	0.320	0.186	-4.098	**0.000**				
	R3off	2.538	0.347							
Pair 4	**R4on**	2.528	0.342	0.000			-1.995			**0.058**
	R4off	2.579	0.371							
Pair 5	**R5on**	2.545	0.363	0.000			2.795			**0.059**
	R5off	2.563	0.388							
Pair 6	**R6on**	2.528	0.342	0.093	-4.034	**0.000**				
	R6off	2.593	0.366							
Pair 7	**R7on**	2.527	0.354	0.167	-2.965	**0.006**				
	R7off	2.584	0.389							
Pair 8	**R8on**	2.532	0.358	0.218	-2.311	**0.027**				
	R8off	2.580	0.398							
Pair 9	**R9on**	2.517	0.356	0.003			0.477	-1.573	**0.116**	
	R9off	2.535	0.391							
Pair 10	**R10on**	2.524	0.357	0.522	-0.436	**0.666**				
	R10off	2.532	0.386							

The positive *z*-score of the Wilcoxson test for the Run 1 indicates larger pupil diameter in the OnRun period. Although, *t*-values of the dependent T-Test and mean values from the descriptive statistics show that the average diameter of the pupil was larger during the OffRun period than the OnRun period for all runs except Run 1, these differences are statistically significant only for 4 runs.

From Table 14, we can study that the size of the pupil is larger during the OffRun period than the OnRun period. However, after conducting statistical tests, we cannot conclude that higher cognitive load leads to a decrement of the pupil diameter on condition Large display.

Small display

For the condition Small display, we followed the same procedure as for the condition Large display. Table 15 shows the results of the statistical tests. Except Run 4, Run 7 and Run 10, the rest of the data are normally distributed which allows us to conduct dependent T-Test. For the Run 10, we conducted Wilcoxson signed-rank test, since the data is symmetrical in shape ($skewness = -0.127$). For the Run 4 and Run 7, the Sign test was conducted.

Except the Run 1, the results of the tests are statistically significant which means that the average diameter of the pupil recorded in the OnRun and OffRun periods differ significantly. The positive *t*-value of the dependent T-Test for the Run 1 indicates larger pupil diameter for the OnRun period. However, negative *t*-values, negative *z*-score and mean values of the descriptive statistics show that for the rest 9 runs the average diameter of the pupil for the OffRun period is higher than the average diameter of the pupil for the OnRun period and these differences are statistically significant for all 9 runs.

Studying Table 15, we can conclude that in contrast to the condition Large display, in condition Small display the pupil dilation is an indicator to the level of cognition and higher cognitive load leads to a decrement of the pupil diameter.

In the condition Large display, we could not conclude that higher cognitive load leads to a decrement of the pupil diameter, however, from the descriptive statistics, we saw that pupil diameter recorded during the OnRun period (higher cognitive load) is lower than the OffRun period (lower cognitive load).

Table 15. Results of the dependent T-Test, Wilcoxson signed rank-test and Sign test for the gaze event Pupil dilation, condition Large display. Normality is checked with the Shapiro-Wilk test.

Pupil dilation -Small display		Descriptive Statistic		Shapiro -Wilk Test	Dependent T-Test		Symmet -ricity	Wilcoxson Test		Sign Test
		Mean diam. pupil	Std. Deviation	p-value	t-value	p-value	Skewness	z-score	p-value	p-value
Pair 1	**R1on**	2.816	0.428	0.818	1.008	**0.321**				
	R1off	2.788	0.483							
Pair 2	**R2on**	2.688	0.437	0.991	-2.863	**0.007**				
	R2off	2.740	0.449							
Pair 3	**R3on**	2.716	0.414	0.082	-2.383	**0.023**				
	R3off	2.760	0.437							
Pair 4	**R4on**	2.680	0.396	0.003			-1.491			**0.024**
	R4off	2.751	0.431							
Pair 5	**R5on**	2.707	0.419	0.391	-2.698	**0.011**				
	R5off	2.749	0.453							
Pair 6	**R6on**	2.711	0.419	0.395	-4.532	**0.000**				
	R6off	2.777	0.446							
Pair 7	**R7on**	2.694	0.414	0.004			0.929			**0.001**
	R7off	2.750	0.462							
Pair 8	**R8on**	2.702	0.420	0.468	-2.802	**0.008**				
	R8off	2.750	0.442							
Pair 9	**R9on**	2.677	0.404	0.211	-2.679	**0.011**				
	R9off	2.720	0.432							
Pair 10	**R10on**	2.688	0.418	0.025			-0.127	-2.992	**0.003**	
	R10off	2.7355	0.458							

It is worth to mention that only for the Run 1 the average pupil diameter for the OnRun period is larger than the OffRun period on both conditions. In the beginning of the task, pupils are very sensitive to the light produced on the displays. Due to this pupillary

response in the first run the size of the pupil is higher. For later runs, pupils adjust to the light of the display.

From the results of the condition Small display, we have learned that the differences between the data collected during the OnRun period (higher cognitive load) and OffRun period (lower cognitive load) are statistically significant and the average diameter of the pupils is higher on the OffRun period. For the condition Large display, we could not find the statistical differences between these two periods, however, using the mean values from the descriptive statistics, we can argue that higher cognitive load leads to a decrement of the pupil diameter. Using this knowledge in the next section, we will conduct the statistical analysis to understand how the pupil dilation changes between the large and small displays.

Large vs. Small display

As we already discussed, pupil dilation can be used as an indicator of the level of cognition. And we concluded that higher cognitive load leads to a decrement of the pupil diameter. Using this knowledge, we can conduct another test to understand if there is a difference in cognitive load while participant performing the task on the large and small displays.

In this regard, we will compare the mean values of the pupil diameter between the large and small displays. For comparison, we will use the data recorded only in the OnRun period.

Table 16 shows the results of the comparison between two conditions. For all 10 runs the data is normally distributed and we are allowed to conduct dependent T-Test. The results of the test are statistically significant in every run. It means that the average pupil diameter recorded in two different conditions differ significantly. The negative *t*-values of the dependent T-Test indicate the larger pupil diameter for the condition Small display.

Using gaze event pupil dilation, we found the statistically significant difference between our two conditions on a different level of cognition (for every run of the task). Respectively, display size has an influence on the cognitive load at the run level. Therefore, we do not see the necessity to conduct the statistical test to compare the average pupil diameter for the whole task between our two conditions.

Table 16. Results of the dependent T-Test. The gaze event Pupil dilation, Large vs. Small display comparison. Normality is checked with the Shapiro-Wilk test.

Pupil dilation -Large vs. Small Display		Descriptive Statistic		Shapiro-Wilk Test	Dependent T-Test	
		Mean diam. of pupil	Std. Deviation	p-value	t-value	p-value
Pair 1	**R1L**	2.593	0.361	0.250	-6.633	**0.000**
	R1S	2.816	0.428			
Pair 2	**R2L**	2.483	0.348	0.171	-6.311	**0.000**
	R2S	2.688	0.437			
Pair 3	**R3L**	2.484	0.320	0.094	-7.778	**0.000**
	R3S	2.716	0.414			
Pair 4	**R4L**	2.528	0.342	0.240	-6.541	**0.000**
	R4S	2.680	0.393			
Pair 5	**R5L**	2.545	0.363	0.273	-6.248	**0.000**
	R5S	2.707	0.419			
Pair 6	**R6L**	2.528	0.342	0.324	-6.782	**0.000**
	R6S	2.711	0.419			
Pair 7	**R7L**	2.527	0.354	0.189	-6.889	**0.000**
	R7S	2.694	0.414			
Pair 8	**R8L**	2.532	0.358	0.901	-5.984	**0.000**
	R8S	2.702	0.420			
Pair 9	**R9L**	2.517	0.356	0.881	-6.060	**0.000**
	R9S	2.677	0.404			
Pair 10	**R10L**	2.524	0.357	0.993	-6.843	**0.000**
	R10S	2.688	0.418			

8.5 NASA TLX and the level of performance

In this section, we will discuss the results of Nasa TLX questionnaire and the performance of the participants. The idea behind using NASA TLX questionnaire is to assess the influence of the display size on the cognitive load more in a subjective way. Additionally, using the number of correct answer given by participants as measurements will help us to investigate the influence of the display size on the level of cognition.

In the next chapter, we will discuss how the objective results gained with the help of eye-tracking technology correlate with the results of the NASA TLX questionnaire and the performance of the participants.

Nasa TLX

NASA TLX questionnaire is a rating procedure to rate overall workload score [30]. The questionnaire includes six subjects: *mental demand, physical demand, temporal demand, performance, effort* and *frustration*. Each subject has 20 scales to rate. Each scale is rated with 5 points. The minimum point that each subject can be ranked is 5, the maximum is 100.

Table 17 shows the results of the comparison between the large and small displays. The data is normally distributed only for the subject *Performance* and we conducted dependent T-Test for this data. For the rest of the subjects, we conducted Sign test, since the data did not fulfill the requirements of the Wilcoxson signed-rank test as well.

From Table 17, we can observe that the results of the tests are not statistically significant. However, from the mean values of the descriptive statistics, we can see that except the subject *Performance*, for the rest of the subjects, condition Large display has higher points than the condition Small display. However, none of these differences are statistically significant. The negative *t*-value of the dependent T-Test for the subject *Performance* shows that participants rated their performance higher on the small display, however, this result is statistically not significant either.

Table 17. Results of the dependent T-Test and Sign test for the dependent variable Nasa TLX, Large vs. Small display comparison. Normality is checked with the Shapiro-Wilk test.

Nasa TLX -Large vs. Small Display		Descriptive Statistic		Shapiro -Wilk Test	Dependent T-Test		Symmet -ricity	Wilcoxson Test		Sign Test
		Average rank	Std. Deviation	p-value	t-value	p-value	Skewness	z-score	p-value	p-value
Mental demand	**Large**	13.353	5.039	0.007			1.189			**0.701**
	Small	12.500	4.568							
Physical demand	**Large**	7.265	5.744	0.000			1.861			**0.832**
	Small	6.382	5.211							
Temporal demand	**Large**	13.471	4.350	0.048			0.686			**0.362**
	Small	12.324	4.036							
Performance	**Large**	10.147	3.332	0.173	-1.011	**0.320**				
	Small	10.971	3.973							
Effort	**Large**	14.471	3760	0.005			1.291			**0.860**
	Small	13.765	4.171							
Frustration	**Large**	9.735	5.976	0.006			0.680			**0.832**
	Small	9.529	5.780							

Performance

As we discussed in the experimental design, implemented visual search task creates the log file at the end of the task. This file contains the real time of the clicks on the target item. Using this information, we counted the number of correct clicks for all participants on the large and small display. In this section, we will compare the performance of the participants while performing the task on the large and small displays.

Table 18 shows the results of the statistical tests for the number of correct answers. Except Run 9, the data is not normally distributed. For the Run 9, we conducted dependent T-Test. The data from the Run 5, Run 7 and Run 10 are moderately skew, which means the non-parametric Sign test can be conducted. For the rest of the runs, we conducted Wilcoxson signed-rank test.

Table 18. Results of the dependent T-Test, Wilcoxson test and Sign test for the dependent variable Performance, Large vs. Small display comparison. Normality is checked with the Shapiro-Wilk test.

Nº Correct Answers Large vs. Small displays		Descriptive Statistic		Shapiro-Wilk Test	Dependent T-Test		Symmet-ricity	Wilcoxson Test		Sign Test
		Nº Correct answers	Std. Deviation	p-value	t-value	p-value	Skewness	z-score	p-value	p-value
Pair 1	**R1Large**	2.941	0.239	0.000			0.309	0.277	**0.782**	
	R1Small	2.912	0.288							
Pair 2	**R2Large**	2.618	0.604	0.000			0.186	-0.806	**0.420**	
	R2Small	2.735	0.448							
Pair 3	**R3Large**	2.118	0.978	0.005			-0.098	-0.653	**0.513**	
	R3Small	2.235	0.781							
Pair 4	**R4Large**	1.735	0.898	0.016			-0.276	-1.125	**0.260**	
	R4Small	1.971	0.834							
Pair 5	**R5Large**	1.412	0.925	0.003			0.628			**0.383**
	R5Small	1.441	0.927							
Pair 6	**R6Large**	1.147	0.784	0.027			-0.153	-2.225	**0.026**	
	R6Small	1.706	1.001							
Pair 7	**R7Large**	1.059	0.952	0.003			-0.646			**0.523**
	R7Small	1.088	0.866							
Pair 8	**R8Large**	1	0.739	0.004			-0.375	0.900	**0.368**	
	R8Small	0.824	0.797							
Pair 9	**R9Large**	0.882	0.913	0.052	0.423	**0.675**				
	R9Small	0.794	0.914							
Pair10	**R10Large**	0.794	0.729	0.006			-0.605			**1.000**
	R10Small	0.853	0.821							

Except the Run 6, the results of the tests are not statistically significant. Additionally, from the mean values of the descriptive statistics, we can observe that for the Run 1, Run 8 and Run 9 the number of correct answers given on the large display is higher. For the rest 7 runs,

the number of correct answers given on the small display is higher. However, these differences are statistically significant only for the Run 6.

9. Discussion

In this chapter, we will discuss the results presented in the previous chapter. In the results chapter, we conducted statistical tests to understand the effect of the cognitive load on each eye tracking measurement and investigated the influence of the display size on each eye tracking measurement which have direct relation to the cognitive load. Thus, we investigated the influence of the display size on the cognitive load.

We will discuss how each gaze event is applicable to answer our main research questions:

How does the display size influence cognitive load in visual search tasks?

Additionally, the correlation between the results of the objective and subjective data analysis will be presented.

9.1 Fixation

With the statistical tests conducted for the gaze event fixation, we can answer to our 1st research sub-question: *How do changes in Cognitive Load influence fixation in Large and Small displays?*

In the 1st phase of the analysis, the results of the tests were statistically significant for both conditions. And we found that higher cognitive load leads to a higher number of fixations. This finding is in line with the previous work done by Barreras [10].

However, in the second phase of our analysis, we did not find any statistically significant differences between the number of fixation of the large and small displays. First, we conducted a test to compare the number of fixation on each run. Only in two runs the results were statistically significant, thus, the large display was cognitively demanding. Additionally, we conducted the statistical test not on each run but for the whole task. Again, we did not find any statistically significant differences between our two conditions.

Figure 22 illustrates the average number of fixations recorded during the active search on every run. Observing the figure, we can see that except the 1st, 9th, and 10th run, the number of fixations recorded on the large display is slightly higher.

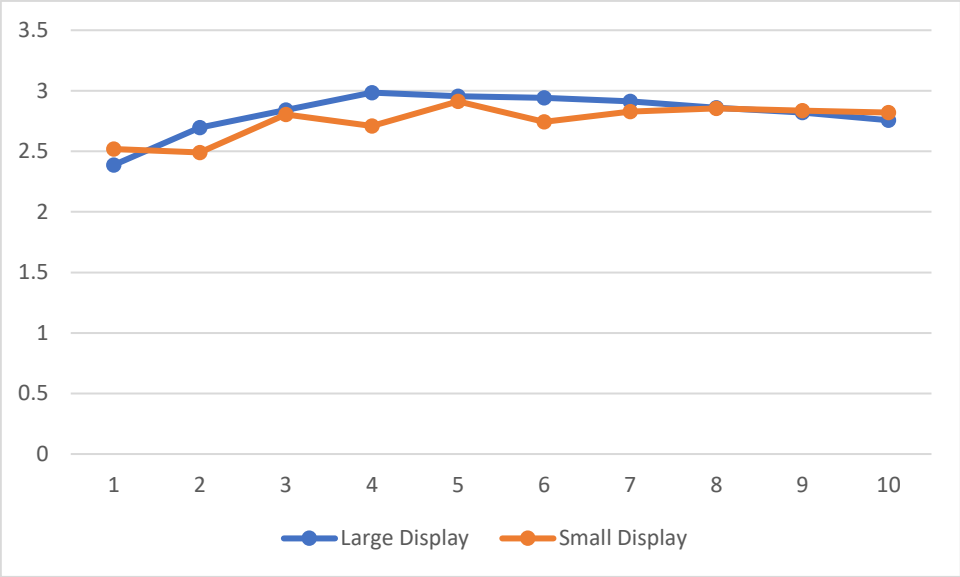


Figure 22. Comparison between the mean values of the number of fixations recorded for each run on the large and small displays. Data recorded during OnRun period (active search).

Additionally, Figure, 23 shows the average number of fixations recorded during the OnRun period for the whole task. As we can see, performing the task on the large display produced slightly higher number of fixations which is the indicator of the higher cognitive state.

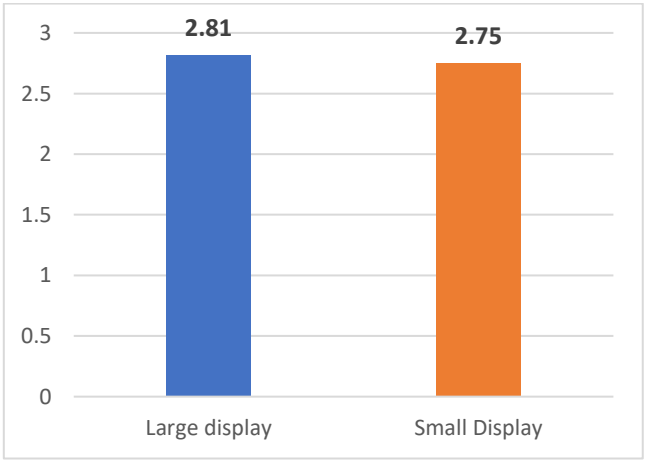


Figure 23. Average number of fixations recorded during the active search (OnRun) for the whole task.

However, based on the statistical analysis, we did not find statistically significant differences between our conditions using the measure fixation. Analysing mean values shows that there is a tendency for the large display to lead higher cognitive load, however, we will conclude that, the display size has almost no influence on the cognitive load in visual search task.

9.2 Saccade

In this section, we can answer to our 2nd research sub-question based to our statistical analysis: *How do changes in Cognitive Load influence saccade in Large and Small displays?*

In the 1st phase of the analysis, the results of the tests were statistically significant for both conditions. And we found that higher cognitive load leads to a higher number of saccades. This finding is in line with the previous work done by Barreras [10].

In the second phase of the analysis, we conducted the tests to compare the mean number of saccades between the large and small displays. Statistical tests conducted for each run of the tests shows that there are not statistically significant differences between our conditions except Run 4. Additionally, we conducted another test to compare the number of saccades for the whole task and we found statistically significant differences between the large and small displays.

Figure 24 illustrates the average number of saccades recorded on the large and small displays. From the figure, we can observe that except the first and last runs, the number of saccades produced on the large display is higher.

Additionally, from the Figure 25, one can see that the average number of saccades recorded for the whole task is higher on condition Large display.

As for the measurement fixation, for the measurement saccade we can see that there is a tendency for the large display leading to a higher cognitive load. Additionally, in the previous chapter, we found statistically significant difference between the large and small displays. Thus, performing the on the large display produces higher number of saccades.

Based on our statistical analysis, we can conclude that large display leading to a higher cognitive load.

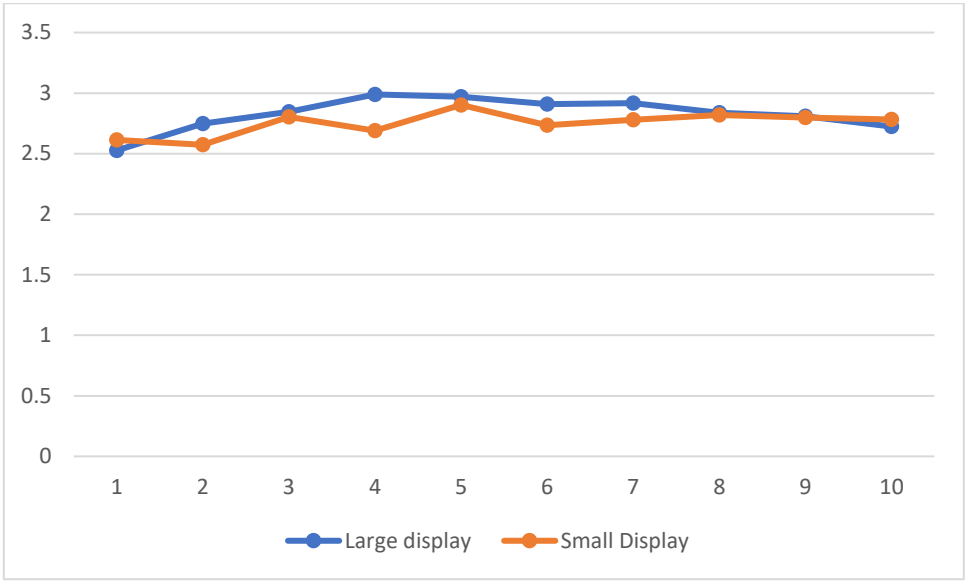


Figure 24. Comparison between the mean values of the number of saccades recorded for each run on the large and small displays. Data recorded during OnRun period (active search).

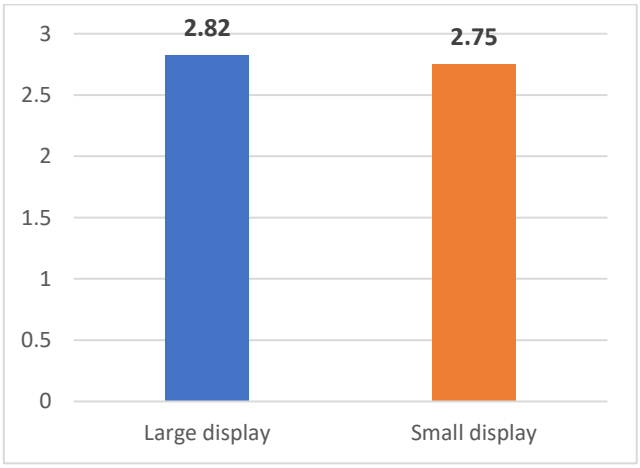


Figure 25. Average number of fixations recorded during the active search (OnRun) for the whole task.

9.3 Blink

Based to the statistical analysis, we will answer to our 3rd research sub-question: *How do changes in Cognitive Load influence blinks in Large and Small displays?*

In the 1st phase of the analysis, the results of the tests for the condition Large display were statistically significant for all runs. For the condition Small display except one run, we also found the statistically significant differences. And we concluded that higher cognitive load leads to a lower number of blinks. This finding is in line with the previous work done by Barreras [10].

In the second phase of the analysis, we conducted the tests to compare the mean number of blinks between the large and small displays. First, we conducted statistical tests for each run. Except Run 7, we did not find statistically significant differences between our two conditions. Additionally, we conducted another test to compare the number of blinks for the whole task and we found statistically significant differences between the large and small displays.

Figure 26 illustrates the average number of blinks recorded during the active search on every run. Observing the figure, we can see that except the 6th and 9th run, the number of blinks recorded on the small display is higher.

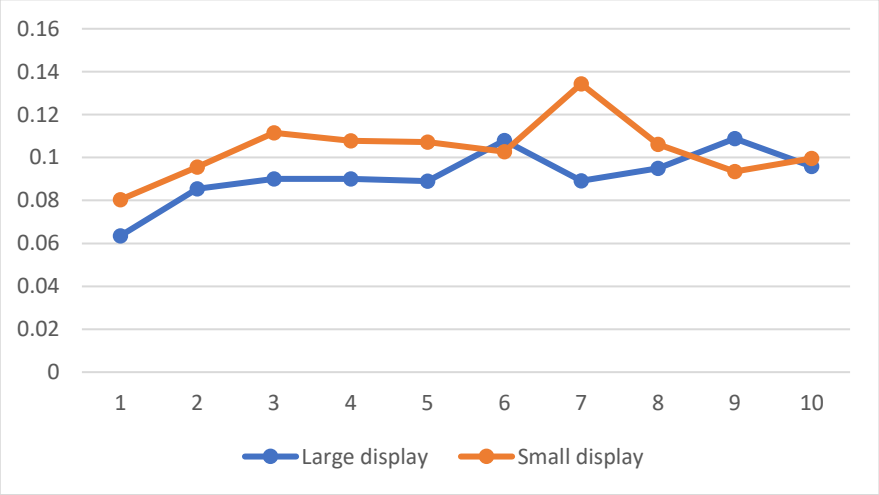


Figure 26. Comparison between the mean values of the number of blinks recorded for each run on the large and small displays. Data recorded during OnRun period (active search).

Additionally, from Figure 27, we can observe that the average number of blinks recorded for the whole task is significantly higher for the condition Small display.

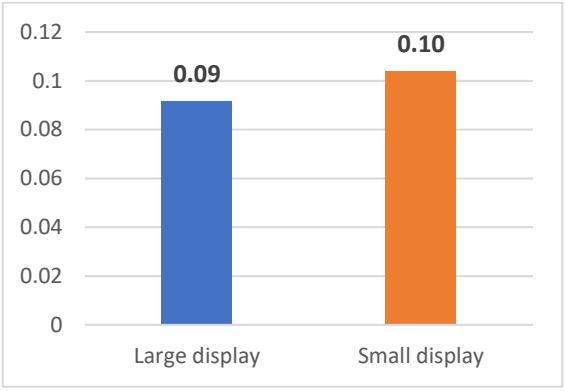


Figure 27. Average number of blinks recorded during the active search (OnRun) for the whole task.

Based to the results, we can conclude that performing the task on the large display is more cognitively demanding.

9.4 Pupil dilation

Pupil dilation is the most investigated gaze event which has a direct relation to the cognitive load. And many researches show that higher cognitive load leads to an increment of the pupil diameter [10][20]. However, our findings with this gaze event are in contrast with other researches.

Based to the statistical analysis, we will answer to our 4th research sub-question: *How do changes in Cognitive Load influence blinks in Large and Small displays?*

In the first phase of the analysis, the results of the test for condition Small display were statistically significant, except the 1st run. And the average diameter of the pupil recorded for the OnRun period (higher cognitive load) was significantly lower than OffRun (lower cognitive load) period. This finding is in contrast with previous researches investigated the relation between pupil dilation and cognitive load.

For the large display, however, only in 4 runs we found statistically significant differences. Additionally, from the Figure 28, we can observe that for most of the runs the average pupil diameter for the OffRun period is higher. Therefore, we can conclude that in frame of our research, higher cognitive load leads to a decrement of the pupil diameter.

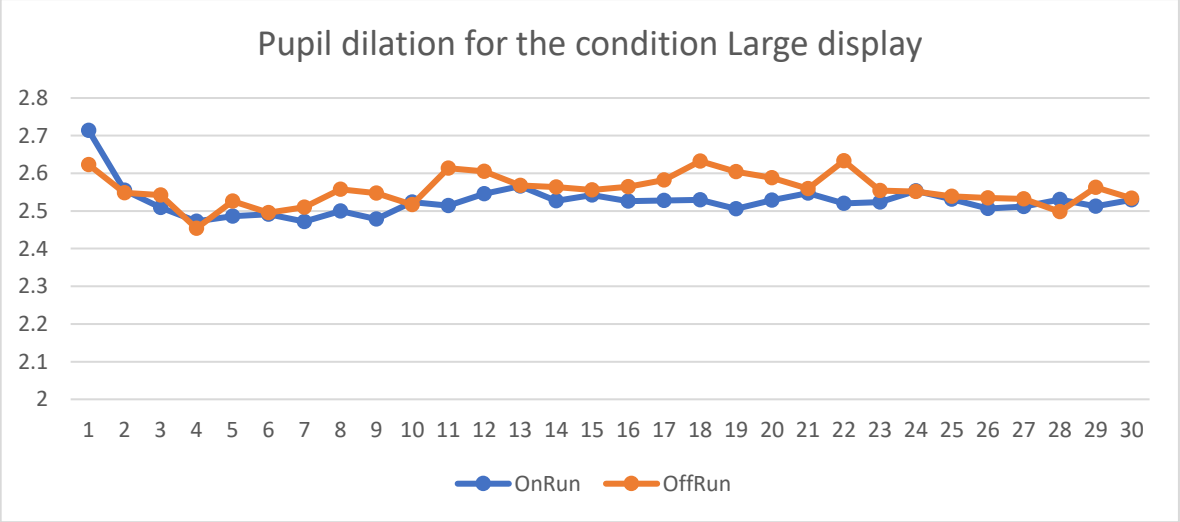


Figure 28. Comparison between the mean values of the pupil diameter recorded for each lap on the large display.

In the second phase of the analysis, we conducted the statistical test to compare the average pupil diameter recorded on the large and small displays. The results of the statistical tests were significant for every run. And we concluded that the pupil diameter of the participants was larger on the condition Small display.

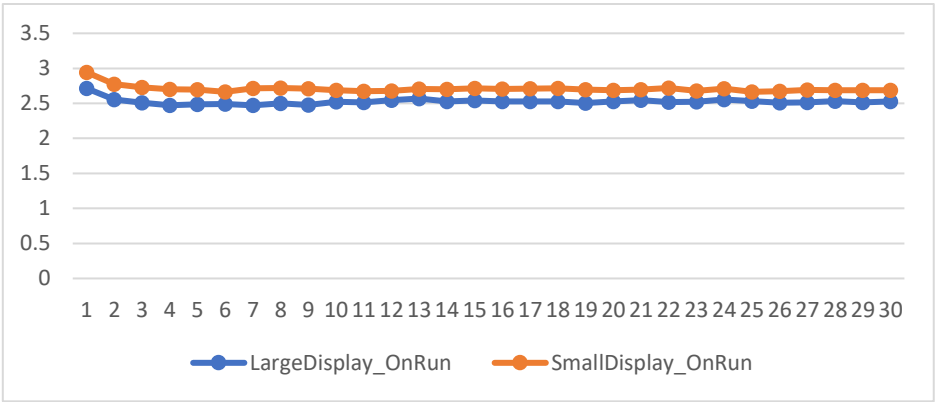


Figure 29. Comparison between the mean values of the pupil diameter recorded for each lap on the large and small displays. Data recorded during OnRun period (active search).

Additionally, from Figure 29, one can observe that the average pupil diameter for the condition Large display is lower for every lap of the task.

As in the first phase of the analysis, we concluded that higher cognitive load leads to a decrement of the pupil diameter in the visual search task, for the second phase of the analysis we can say that performing the task on the large display produces higher cognitive load.

Even though the finding in the first phase of analysis are contrast with the previous works, the finding in the second phase of analysis supports our results of the other gaze events.

Therefore, using pupil dilation, we can conclude that the large display produces higher cognitive load in visual search task.

9.5 NASA TLX and performance

In this subsection, we will answer to our 5th research sub-question: *How do the results of NASA TLX questionnaire and the users' performance correlate with the results gained with the help of eye-tracking technology?*

As we described in the previous chapter, we did not find any statistically significant differences between the subjective data of the large and small displays. However, will not rely on the strict statistical analysis as for gaze event. For the results of the NASA TLX questionnaire and the number of right answers we will discuss the differences of the mean values.

From Figure 30, we can see that except the subject Performance, for the rest subjects the mean values are higher for condition Large display. On the subject Performance, participants rated the small display higher than the large display.

Observing the result of the NASA TLX questionnaire, we can conclude that the large display required more mental and physical effort to perform the task which is in line with our finding using the gaze events.

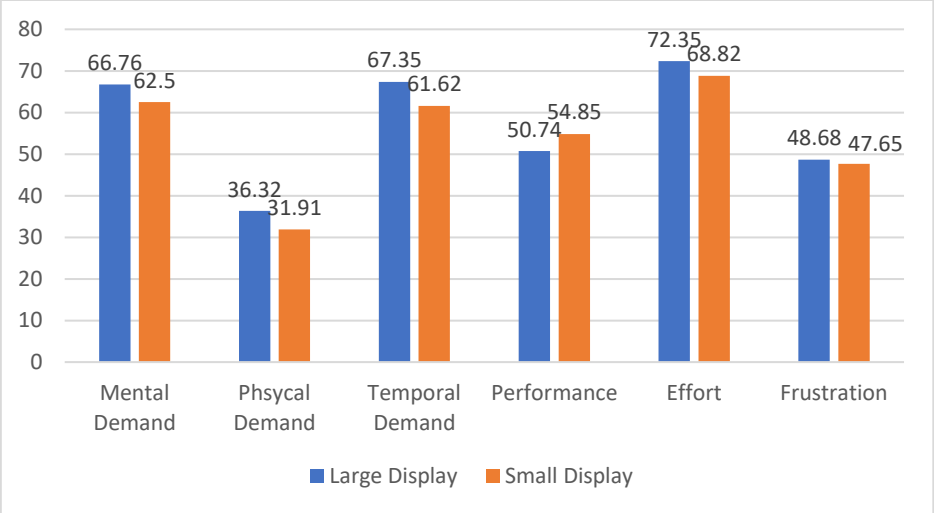


Figure 30. The results of the NASA TLX questionnaire.

To assess the influence of the display size on the performance of participants, we will follow the same procedure. In the previous chapter, we did not find statistically significant differences between the number of right answers of the large and small displays. Therefore, we will discuss the mean values of the number of right answers.

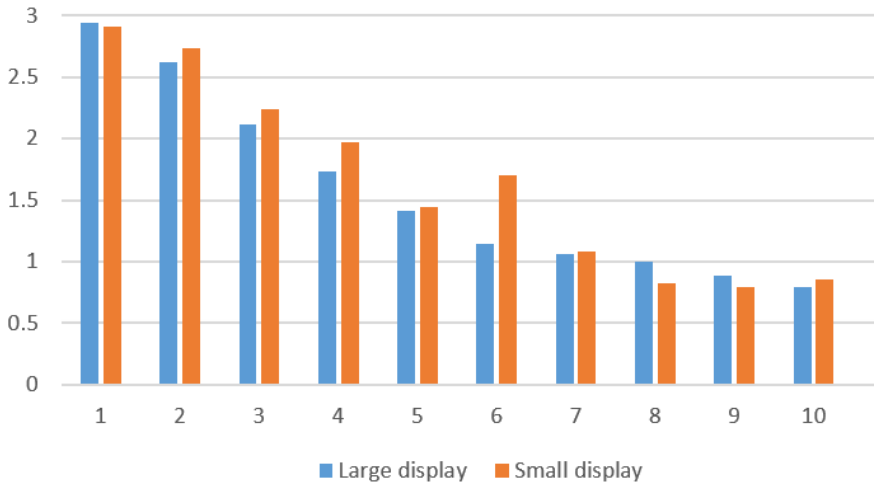


Figure 31. The mean numbers of the right answers for each run.

As we can see from Figure 31, the number of right answers for both conditions are decreasing while increasing the task complexity and except the 8th and 9th run, the number of right answers are higher on the condition Small display.

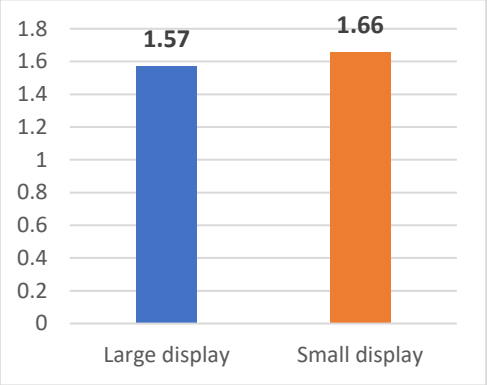


Figure 32 Average number of right answers given on each condition for the whole task.

Additionally, from Figure 32, we can see that the average number of correct answers given on the condition Small display is slightly higher.

Although, we did not find statistically significant differences between our two conditions using NASA TLX questionnaire and the level of performance, we can still conclude that there is a tendency for the large display leading to a higher cognitive load.

9.6 Summary

In this section, we concluded that cognitive load has a direct influence on each gaze movement in the visual search task. Furthermore, changes on the level of cognition leads to the changes on the gaze measurements in the same way for both – small and large displays.

Conducting the analysis, we were able to answer to our main research questions:

How does the display size influence cognitive load in visual search tasks?

We answered to our research question using gaze measurements, NASA TLX questionnaire and the performance rate of the participants.

Using the metric fixation, we showed that display size does not influence the cognitive load in visual search task and performing the task on the large and small displays yields almost the same cognitive load. However, using the gaze measurements saccade, blinks and pupil dilation, we showed that display size has a little influence on the cognitive load, thus, performing the task on the large display tends to produce higher cognitive load.

Additionally, the result of the analysis of performance rate and NASA TLX questionnaire are in line with the aforementioned findings.

In our research, we found unexpected results related to the pupil dilation. However, with the help of pupil dilation, we were able to answer to our main research question. And the results for this gaze event was in line with the results of other dependent variables.

Furthermore, in the end of our experiment, we asked a direct question to the participants in the post-questionnaire (see Appendix A): *Which screen did you find most difficult to work with?*

18 out of 34 participants found the large display and 15 participants the small display difficult to work with. One participant believed that display size did not affect her performance.

10. Conclusion and future work

In this thesis, we conducted an empirical analysis to understand how the display size influence the cognitive load in visual search tasks. Through the literature review, we have learned that the level of cognitive state has a direct influence on the eye movements and using eye-tracking technology the level of cognition can be estimated. Therefore, we also used the eye tracking technology as a tool to measure the cognitive state.

After the first phase of our analysis, we have also concluded that it is possible to measure cognitive load in visual search tasks using measurements such as fixation, saccade and blinks and these findings are in line with the findings of the research conducted by Barreras [10]. Additionally, we showed that these findings are valid not for one display, but for different sized displays.

However, for the gaze event pupil dilation, our findings are in contrast with the previous findings on the relation the level of cognition and the pupillary response. One possible explanation for this could be the different way of analysing the data. We used statistical tests to analyse the pupil dilation, namely, we compared the average pupil diameter collected on the high and low level of the cognitive states. However, in the previous work, Barreras [10] analysed the pupil dilation using not statistical tests, but the time series analysis.

Although, our result for the gaze event pupil dilation does not match with the previous works, we have found significant differences between the data collected during the higher and lower cognitive load.

After finishing the first phase of our analysis, we answered to each research sub-questions and we have learned how the changes in cognitive load influence each gaze measurement on the large and small displays.

After finishing the second phase of our analysis, we were able to answer to our main research question and we concluded that large display leads to a slightly higher cognitive load in visual search tasks.

Additionally, using gaze measurements fixation, saccade and blinks, we showed that difference between the cognitive load for two displays are not detectable for short period of time. Thus, at the run level, we did not find any statistically significant differences between our two conditions. However, the difference was detectable only at the task level (longer period of time) and using these gaze measurements, we showed that there is a tendency for the large display leading to a slightly higher cognitive load.

As a future work, it would be interesting to conduct an experiment not in the controlled lab environment, but investigate the influence of the display size during the real search activities, e.g. searching the number of the lecture room on the large and small displays or checking the timetable using different sized displays. However, one has to be careful with the measurement pupil dilation while conducting the field study, since lights from the environment can cause a pupillary response.

Additionally, investigating the relation between the cognitive load and other gaze measurements such as fixation duration, saccade velocity and even the complex measurement scanpath can be suggested as a future work.

References

[1] J. Zagermann, U. Pfeil, and H. Reiterer, “Measuring Cognitive Load Using Eye Tracking Technology in Visual Computing,” *Proc. Sixth Work. Beyond Time Errors Nov. Eval. Methods Vis.*, pp. 78–85, 2016.

[2] F. Chen *et al.*, “Robust multimodal cognitive load measurement,” *Robust Multimodal Cogn. Load Meas.*, pp. 199–204, 2016.

[3] E. Control, “Mechanisms of Active Maintenance and Executive Control Edited by,” *System*, vol. 21, pp. 1–27, 1999.

[4] G. Alan D.Baddeley, “Working Memory,” *Psychol. Learn. Motiv.*, vol. 8, pp. 47–89, 1974.

[5] W. W, “Variations in psychology,” *New York Wadsworth*, no. 9 ed., pp. 281–282, 2013.

[6] J. Klingner, “Measuring cognitive load during visual tasks by combining pupillometry and eye tracking,” *Perspective*, no. May, p. 130, 2010.

[7] G. A. Miller, “The magical number seven, plus or minus two: some limits on our capacity to process information,” *Psychol. Rev.*, pp. 81–97, 1956.

[8] da S. F. L. Niedermeyer E., “Electroencephalography: Basic Principles, Clinical Applications, and Related Fields.,” *Lippincott Williams & Wilkins.*, 2004.

[9] O. V. Hämäläinen, Matti; Hari, Riitta; Ilmoniemi, Risto J.; Knuutila, Jukka; Lounasmaa, “Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain,” *Rev. Mod. Physics.*, pp. 413–497, 1993.

[10] E. Barreras Aragón, “Measuring Cognitive Load using Eye-Tracking in visual search tasks,” p. 146, 2017.

[11] H. Y. Chi Jian-nan, Zhang Peng-yi, Zheng Si-yi , Zhang Chuang, “Key Techniques of Eye Gaze Tracking Based on Pupil Corneal Reflection,” *Intell. Syst. 2009. GCIS ’09. WRI Glob. Congr.*, 2009.

[12] “What is eye tracking and how does it work? - iMotions.” [Online]. Available: https://imotions.com/blog/eye-tracking-work/. [Accessed: 06-Jan-2018].

[13] “Glint position changing according to point of regard.” [Online]. Available: https://www.researchgate.net/figure/233969486_fig2_Figure-2-Glint-position-changing-according-to-point-of-regard. [Accessed: 06-Jan-2018].

[14] J. van de W. Kenneth Holmqvist, Marcus Nystrom, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, *Eye Tracking: A Comprehensive Guide to Methods and Measures*. 2012.

[15] C. J. Ellis, “The pupillary light reflex in normal subjects.,” *Br. J. Ophthalmol.*, vol. 65, no. 11, pp. 754–759, 1981.

[16] A. F. Kramer, “Physiological metrics of mental workload: A review of recent progress.,” *London Taylor Fr. Ltd.*, 1991.

[17] E. Granholm, R. F. Asarnow, A. J. Sarkin, and K. L. Dykes, “Pupillary responses index cognitive resource limitations,” *Psychophysiology*, vol. 33, no. 4. pp. 457–461, 1996.

[18] V. Peysakhovicha, F. Dehaisa, and M. Causseab, “Pupil diameter as a measure of cognitive load during auditory-visual interference in a simple piloting task,” 2015, pp. 5199–5205.

[19] N. Nourbakhsh, Y. Wang, and F. Chen, “GSR and blink features for cognitive load classification,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8117 LNCS, no. PART 1, pp. 159–166, 2013.

[20] S. Chen, J. Epps, N. Ruiz, and F. Chen, “Eye activity as a measure of human mental effort in HCI,” *Proc. 15th Int. Conf. Intell. user interfaces - IUI ’11*, p. 315, 2011.

[21] A. M. Karam and J. E. Hughes, “A Comparison of the Effects of Mobile Device Display Size and Orientation , and Text Segmentation on Learning , Cognitive Load , and User Perception in a Higher Education Chemistry Course,” 2015.

[22] L. Lischke *et al.*, “Using Space : Effect of Display Size on Users ’ Search Performance,” p. 6, 2015.

[23] Vidhu Jain, “Investigating the Influence of Display Size on Aspects of Spatial Memory Vidhu Jain,” *Master Thesis*, p. 107, 2017.

[24] M. Haralambos and M. Holborn, *Sociology, themes and perspectives*, 3rd editio. London: Collins Educational, 1991.

[25] H. M. H.Michael, “Sociology. Themes and perspectives.,” *London*, no. 3rd edition, p. page 727, 1991.

[26] A. M. Treisman and A. Treisman, “A Feature-Integration Theory of Attention,” *Cognitive Psychology*, vol. 12. pp. 97–136, 1980.

[27] B. McElree and M. Carrasco, “The Temporal Dynamics of Visual Search: Evidence for Parallel Processing in Feature and Conjunction Searches,” *J Exp Psychol Hum Percept Perform*, pp. 1571–1539, 1999.

[28] B. Aragón, “How to measure Cognitive Load using Eye- Tracking in visual search tasks,” 2017.

[29] S. A. McLeod, “Independent, dependent and extraneous variables.” [Online]. Available: www.simplypsychology.org/variables.html.

[30] NASA, “Task Load Index,” 2017.

[31] Perceptive Pixel, “Perceptive Pixel,” pp. 1–28, 2011.

[32] G. Charness, U. Gneezy, and M. A. Kuhn, “Experimental methods: Between-subject and within-subject design,” *J. Econ. Behav. Organ.*, vol. 81, no. 1, pp. 1–8, 2012.

[33] A. Blandford, A. Cox, and P. Cairns, “Controlled experiments,” *Res. Methods Human-Computer Interact.*, pp. 1–16, 2008.

[34] A. Rind, “Some Whys and Hows of Experiments in Human–Computer Interaction,” *Found. Trends® Human–Computer Interact.*, vol. 5, no. 4, pp. 299–373, 2011.

[35] A. Field, “Discovering Statistics Using SPSS,” no. Third edition, p. 857, 2009.

[36] R. Dawson, “How Significant Is A Boxplot Outlier?,” *J. Stat. Educ.*, vol. 19, no. 2, pp. 1–13, 2011.

[37] “7.1.6. What are outliers in the data?” [Online]. Available: http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm. [Accessed: 19-Dec-2017].

[38] “Identifying outliers.”

[39] S. S. Shapiro and M. B. Wilk, “An Analysis of Variance Test for Normality (Complete Samples),” *Biometrika*, vol. 52, no. 3/4, p. 591, 1965.

[40] M.G.Bulber, “Principles of Statistics,” *Technometrics*, vol. 9, no. 4. Dover Publications Inc., New York, p. 698, 1967.

[41] V. M. G. Barrios *et al.*, “AdELE: A Framework for Adaptive E-Learning through Eye Tracking,” *Proc. IKNOW*, vol. 4, no. October, pp. 1–8, 2004.

Appendix A

Welcome

Welcome to the experiment. I appreciate your attendance and your time participating in this study. This experiment is part of my Master Thesis and your contribution is essential to support my work. Before getting started, please read the following introduction.

Objective:

We will use Eye-Tracking glasses to collect your gaze movements, including the size of your pupil. We will do this while you perform the Visual Search task.

Study procedure:

The procedure of the study will be as follow: After signing the declaration of consent, I will ask you to fill a demographic questionnaire. Afterwards, I will give you introduction about the system that you are going to use. If you have any questions, please do not hesitate to ask.

First, we need to calibrate the eye-tracker. To do so, I will ask you to look at three specific points. Once the eye-tracker calibrated, you will perform *test run* before starting the real task.

Experiment consists of the task, which you will perform on two different monitors. During the task, you will have to find the element with specific shape and color among other elements. The task has 10 runs. After each run the complexity of the task will increase. If you can find the right answer and click on it, you will hear beep sound. After each run, there will be a circle in the middle of the screen and you should place the mouse pointer inside it.

After completing the task on each monitor, I will ask you to fill NASA TLX questionnaire.

At the end of the experiment, you will have to fill one last questionnaire, and I will ask you some open-ended questions.

Time frame and compensation:

Completing the whole experiment has a maximum duration of 60 minutes. If you feel uncomfortable, you can cancel the experiment at any point in time. Please just notify the study adviser.

After completion of the experiment, you will receive a compensation for your help of 8 Euros.

Finally, I would like to thank you for your participation!

Declaration of consent

ID:_____

Information to study management:

Study advisor: Javid Guliyev

Institution: Human Computer Interaction, University Konstanz

Study procedure:

I would kindly point your attention on the subsequent study procedure: You can cancel the study at any point in time. If you need a break, please feel free to ask for one. If you have any question regarding the basic/general procedure or the system, I am pleased to answer them. However, I ask for your understanding that I cannot answer specific questions about ongoing exercises to prevent biases in the results. After completion of the study, I am happy to answer any of your question.

Declaration:

I was informed about the purpose, content and duration of this study. Within the scope of this study personal data is collected using questionnaires. Additionally, data related to my eyes will be recorded.

I hereby acknowledge that this data will be anonymized, treated with caution and will not be passed to third parties. Data will completely be used for aforementioned purposes and – with my consent – for international presentations.

I hereby declare approval with the above-mentioned points:

Name: _____ Date, place: _____

Signature: _____

Hereby, the study advisor declare that they will use the Eye-Tracking data, as well as any other collected data, completely for research purposes within the framework of the study.

Name: Javid Guliyev Date, place: _____

Signature: _____

ID:_____

Demographic questionnaire

Personal Information:

Gender

- Male
- Female

Age: _____Years old

Height: _____cm

Profession

- Student. Field of study: _____
  - Bachelor
  - Master
  - PhD
- Other: _____

Do you wear Glasses/Contact Lenses?

- Yes, Glasses
- Yes, Contact Lenses
- No

If you have corrective lenses, please indicate your graduation

- Left Eye: _____
- Right Eye: _____

Are you color-blind?

- Yes
- No

Have you ever used an Eye-Tracker before?

- No,
- Yes, once
- Yes, more than once

How often do you use a computer with mouse?

- Everyday
- Several times a week
- Several times a month
- Never

What size of screens do you use in your daily life?

- 7.9 - 12.9 inches
- 13 – 15.6 inches
- 15.6 – 18.5 inches
- 18.5 – 23 inches
- 23 – 27 inches
- 27 inches and bigger

Do you know the resolutions of your screens? If yes, please mention it below:

_____

_____

What kind of tasks do you perform on each of your screens?

_____

_____

_____

_____

_____

NASA Task Load Index

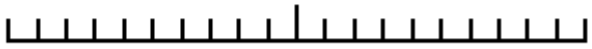
Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?

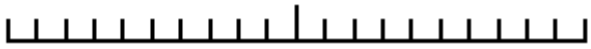
Very Low — Very High

Physical Demand — How physically demanding was the task?



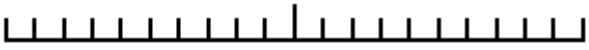
Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?



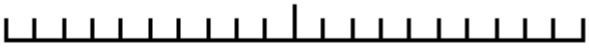
Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?



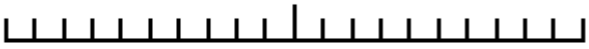
Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?



Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?



Very Low — Very High

ID: _____

Post-questionnaire

Which screen did you find most difficult to work with?

- **Big screen**
- **Small screen**

Can you explain why? Or why the other screen is easier to work with?

_____

_____

_____

I felt comfortable wearing the Eye-Tracker glasses

Strongly disagree | O O O O O | Strongly agree

Using the Eye-Tracker glasses have influenced my performance

Strongly disagree | O O O O O | Strongly agree

Presence of the study advisor has influenced my performance

Strongly disagree | O O O O O | Strongly agree

Appendix B

Content of the USB flash drive

- This thesis as pdf file
- The visual search task
- Programmed Java project for the data preparation
- Log files generated by the visual search task and log files after each transformation phase
- The raw data exported using eye-tracking software and gaze measurement after each step of data preparation
- Demographic information in Excel file
- Documents which were used during the experiment