Measuring Cognitive Load using Eye-Tracking in visual search tasks

Master thesis for the degree

Master of Science (M.Sc.) in Information Engineering

by

Barreras Aragón, Elena

At the

Universität Konstanz

Department of Computer and Information Science

1st Referee: Prof. Dr. Harald Reiterer

2nd Referee: Prof. Dr. Bela Gipp

Konstanz, April 2017

To my parents and
friends for their
unconditional support.

Abstract

The evolution of technology into the era of ubiquitous computing, where computers' actions are mostly based on information about the human state (cognitive, affective and motivational states) is a motivation to design interfaces which provide as much real-time information about the human state as possible. In this regard, it is relevant to study how cognitive load can be applied in the field of Human-Computer Interaction. Over the past few years, the rise of eye tracker manufacturers has produced a diverse range of analysis software and scientific tools to investigate how the human gaze can provide information about the level of cognition based on single gaze events such as fixations or pupillary response. There have been several studies in the fields of eye tracking and cognition states based on visual and auditory tasks; e-learning, visualization of interfaces, etc. But little has been investigated in relation to cognitive load and visual search tasks.

This master thesis deals with the topic of eye tracking as a tool to measure cognitive load in visual search tasks. The design and conduction of an experiment, the posterior analysis based on the *combination* of gaze events; and the results obtained, are essential to get a deeper understanding of the level of cognition presented in visual search tasks, in contrast to researches where only a relation between *single* gaze events and cognitive load is provided.

DECLARATION:

I herby affirm that I have independently written the attached Bachelor’s/Master’s thesis on the topic:

____

____

____

and have not used any other aids or sources other than those I have indicated.

For parts that use the wording or meaning coming from other works (including the Internet and other electronic text and data collections), I have identified them in each case by reference to source or the secondary literature.

Furthermore, I hereby affirm that the above mentioned work has not been otherwise submitted as a thesis for a Bachelor’s/Master’s examination *). I further understand the pending completion of the review process I must keep the materials available that can prove this work was written independently.

After the examination process has been completed, the work will be submitted to the Library of the University of Konstanz and catalogued. It will thus be available to the public through viewing and lending. The described data given, such as author, title, etc. are publicly available and may be used by third parties (for example, search engine providers or database operators).

As author of the respective work, I **consent** / **do not consent** to this procedure ***)**.

A current confirmation of my enrolment is attached.

____

(Signature)

____

(Place, date)

*) Please delete as appropriate

Conventions

Throughout this thesis the following conventions are used:

- The plural “we” will be used throughout this thesis instead of the singular “I”, even when referring to work that was primarily or solely done by the author.

- Unidentified third persons are always described in male form. This is only done for purposes of readability.

Index

List of Figures

List of Tables

1. Introduction

Assessing the cognitive load imposed by visual search tasks is essential to the design of cognitively efficient visual interfaces in HCI. Based on the state-of-the-art in the fields of eye tracking and cognitive load, we decide to design, conduct and analyze an experiment in order to study the relation between eye movements and cognitive load by analyzing the mental effort.

Based on this idea, we come up with the following research question:

> *"How to measure cognitive load with eye tracking in visual search tasks?"*

As aforementioned, the two main focus points are eye tracking and cognitive load. Many researchers have already focused their studies in these fields, but none have studied them in depth. Additionally, most researchers focus on the analysis of single measurements through the design of too complex tasks.

In this regard, our contribution with this master thesis is to research about the state-of-the-art of cognitive load in fewer explored fields, such as visual search, focusing on the analysis of several eye tracking measurements, such as fixations, saccades, blinks and pupil dilation.

With the knowledge gained from our research, we intend to design, develop, conduct and analyze an experiment where one can assess the cognitive load by the study of different eye tracking metrics.

The remaining thesis is structured as follows:

In Chapter 2, we present a theoretical background to describe the definition of cognitive load (section 2.1) and different gaze events to assess the level of cognition through the use of an eye tracker (section 2.2).

In Chapter 3, we present a summary of already conducted studies in eye tracking and its applicability to different fields of research in order to assess cognitive load and emotional states. These studies in fields such as visual and auditory tasks, e-learning, visualization of interfaces etc., have served us as inspiration to define our research question.

In Chapter 4, we define our research question motivated by our findings drawn from our literature review. Moreover, we propose five sub-questions which will help us to get a deeper understanding of our research question.

In Chapter 5, we describe the conduction of a pre-study, the definition of the task (section 5.1), the settings of the experiment (section 5.2), the conduction of the experiment (section 5.3), the analysis of the eye tracking data (section 5.4) and the challenges that one has to take into account when designing a proper study (section 5.5).

In Chapter 6, the experimental design summarizes the study description and the setup of the redefined experiment. We present a description of the tasks (section 6.1) and their implementation using web technology (section 6.2). Section 6.3 describes the apparatus used, followed by section 6.4, which presents the recruitment of participants in the experiment. Finally, we present the procedure followed to perform the experiment (section 6.5).

In Chapter 7, the analysis of the collected data is described taking into account some necessary considerations in the data collection and data preparation (section 7.1). Section 7.2 describes the procedure carried out to analyze the eye tracking data and the NASA TLX questionnaires.

In Chapter 8, we present the results obtained through the analysis of the gaze events: fixations (section 8.1), saccades (section 8.2), blinks (section 8.3), pupil dilatation (section 8.4), and the NASA TLX questionnaires (section 8.5) in order to answer our research question and sub-questions.

In Chapter 9, we discuss the results presented throughout the previous chapters, focussing on the applicability of the gaze events to estimate the level of cognition imposed by each of our tasks. In section 9.1 we discuss each research objective separately, while in Section 9.2 we address the overall research question, based on the discussion of the individual research objectives.

Finally, Chapter 10 presents the conclusions obtained and provides recommendations for future work and the conduction of future experiments.

2. Theoretical Background

This thesis deals with new direct and objective methods for measuring cognitive load, such as eye-tracking. In this regard, it is appropriate to present the operational definition of the term *cognitive load* and how it can be assessed through the use of an *eye tracker*.

Therefore, this chapter is divided into two sections. The first one presents the definition of cognitive load, the three different types of cognitive load defined by Sweller (1999) and different approaches to measuring it. The second section presents a description of how the gaze is captured by the eye tracker, which kind of eye trackers are available and a selection of the most relevant eye events the eye tracker can process.

2.1 Cognitive load

Cognitive load has been long used by psychologists as a synonym of "processing load", "task load" or "mental effort" in order to describe active mental states during the process of problem-solving. Its history can be traced to the beginning of Cognitive Science in the 1950s and to G.A. Miller et al. (1960), who suggest that the working memory capacity has inherent limits. Indeed, cognitive load analogies are defined in relation to people's limited capacity for cognitive tasks and the fact that engaging in one mental task interferes with one's ability to be involved in others.

According to the original model of Baddeley & Hitch presented in the publication of Brünken et al. (2003), our brain is divided into two working memory subsystems where the information is stored. The *visuospatial sketchpad* subsystem accumulates all visual and spatial information, such as written text or pictures; while the *phonological loop* subsystem is in charge of laying up phonological information such as spoken text or music. A representation of the working memory system can be seen in Figure 1.

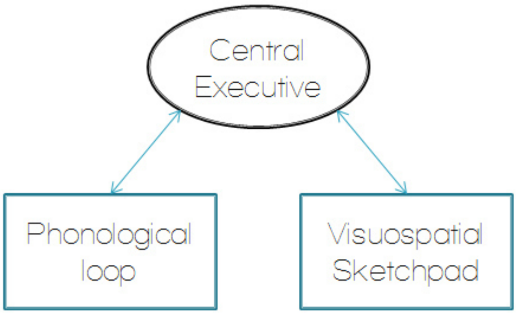


Figure 1. The working memory system.

Both subsystems are independent of one another and both have a limited capacity. In this regard, each of the subsystems presents its own cognitive load. Moreover, the lack of processing capacity of one subsystem cannot be compensated by the other subsystem.

The cognitive load presented in each subsystem is defined by Sweller (1999) as the sum of three kinds of load:

- Intrinsic load: this load is imposed by the structure and complexity of the materials. In this regard, the designer cannot influence or manipulate this kind of load.
- Extraneous load: this load is only caused by the format and the way in which the information is presented and it does not contribute to an understanding of the materials. However, it requires a certain level of working memory and

it can be influenced by the designer. Thit is the reason why in our research, we will focus mainly on this type of cognitive load.

- Germane load: this load is imposed by the effort one needs to make to process the materials in order to understand them.

As aforementioned, each subsystem presents his own and independent level of processing capacity. When the level of cognitive load defined as the sum of intrinsic, extraneous and germane load under particular conditions is higher than the processing capacity of the subsystem, the individual experiments a high cognitive load state or overload.

Brünken et al. (2003) demonstrate this theory by conducting an experiment in which the same information is presented to two individuals through different channels. The first one receives the material as a picture and a text, while the second participant receives the same information as a picture and a narration. Analyzing the results obtained, they conclude that the former participant experiences a higher level of cognition as a consequence of utilizing only the visual memory subsystem (picture and on-screen text). On the contrary, the level of cognition of the latter participant is lower, as a consequence of processing the information separately in each of the subsystems, causing a distribution of the load among the visual and the auditory sub-system.

In this regard, it takes special relevance to be able to calculate the level of cognition, in order to design efficient interfaces that distribute the cognitive load equally in both subsystems. But, how can one measure the cognitive load?

Traditionally, the level of cognition has been measured through techniques such as electroencephalography (EEG) and magnetoencephalography (MEG). These techniques capture changes in magnetic fields at the scalp caused by changing electrical currents in brain neurons, having the advantage of their millisecond-level precision.

Additionally, one can obtain direct and subjective information about the brain activity and cognitive load by the analysis of non-neural techniques such as blood pressure, heart rate, electrical activity in facial muscles, eye movement and pupillary response. We will mainly focus on the last two measurements.

Through the use of an eye tracking, one can gather information about the eye movements and pupil dilation and study their relation to cognitive load. Furthermore, if we combine these direct and objective measurements with indirect and subjective metrics to report the amount of mental effort, such as post-treatment questionnaires like NASA TLX; we can gain a deeper understanding of the level of cognition imposed by a task and apply this knowledge in the context of HCI.

2.2 Eye tracking

In a human eye, three pairs of muscles are responsible for the movements of the eyeball. The Donder's law describes the direction of the gaze as a horizontal, vertical or torsional (roll) movement of the eyeball. The direction of the gaze is uniquely decided by the orientation, independent of how the eye was previously orientated (Holmqvist and Nyström (2011)).

Eye tracking refers to the process of defining the orientation of the gaze by capturing the reflection of the light in the eye. This reflection is recorded by the eye tracker and it is used to describe the path movement. Pupil and corneal reflection are the most dominating eye-tracking methods to capture the gaze.

These eye trackers illuminate the eye with one or more infrared sources, provoking a reflection of the light into the cornea. Due to the concavity and convexity of the lens, the light reflected by the eye is captured by the eye tracker upside down. In this regard, one has to be especially careful when interpreting the corneal reflection. In Figure 2 one can see an example of the corneal reflection's meaning when looking into the camera. If the reflection of the cornea is situated at the bottom, the user is looking directly above the camera. Likewise, if the corneal reflection is situated on the right side, the user is looking to the left of the camera.

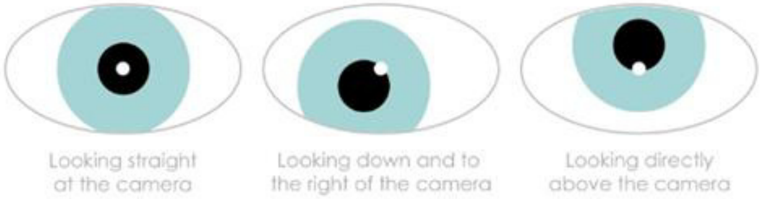

Looking straight at the camera
Looking down and to the right of the camera
Looking directly above the camera

Figure 2. Reflection of the light [Dawson (2015)]

Eye trackers can be classified according to the camera position into *stationary* and *mobile* eye trackers. The former ones can, in turn, be classified as *remote* or *tower-mounted* eye trackers. When the camera is situated below the monitor, they are called *remote* eye trackers and when the eye tracker is fixed in a frame outside the participant's head they are called *tower-mounted* eye trackers. The major advantage of *stationary* eye trackers is the highest accuracy and precision of the data.

Mobile eye trackers allow free movement of the head and can be useful when observing an object that has a three-dimensional structure or when the participant needs to freely move around. In our experiment, we utilize a mobile eye tracker, allowing the participant to feel more comfortable when standing in front of the display.

There are several eye tracking events depending on the position and movement of the gaze, although we describe only the ones that are relevant to measure the level of cognition:

Fixations are the most common eye tracking event. They refer to the period of time where the eye remains still. Fixations are voluntary movements that last from 200-300 milliseconds up to several seconds. The number of fixations indicates the number of times that a participant looked at a certain area of interest (AOI). The duration of a fixation indicates how long a participant looked to a certain AOI. Both measurements are usually related to cognitive load, as we will explain in Chapter 3.

Saccades refer to a shift between two locations, from one fixation point to another. Like fixations, saccades are voluntary movements that take 30 to 80 milliseconds to be completed. They are, in fact, the fastest movement the human body can produce. An interesting measurement from saccades is the number of them. It can indicate cognitive load if the number of saccades is elevated, as discussed in Chapter 3. Additionally, one can measure saccades' velocity and saccades' amplitude, although their relation to cognitive load is not completely clear.

Blinks can be a voluntary movement, although most of the time they are involuntary. Some researchers (García Barrios et al. (2004), Chen et al. (2011), relate the blink rate to states of attention, such as tiredness. Its applicability to measuring cognitive load will be studied in further chapters.

Pupil dilation is an involuntary movement. The pupil reacts to changes in the luminance but also it is reported to react to emotional states and level of cognition. Regarding the former, variations in the brightness of the environment produce the dilation of the pupil (when the environment becomes darker) or its contraction (when the environment becomes brighter). As the luminance is a constant factor, one should control the lab settings to avoid that variations in the brightness can be attributed to changes in the level of cognition. Several types of research have been conducted in this regard and will be explained in the following chapter.

3. Related work

In the last several years, researchers have shown their interest in finding a relation between gaze movements and cognitive load. In the following chapter, we present related work to illustrate which fields have been most extensively explored in relation to cognitive load; especially identifying which eye-tracking measurements best describe cognitive load and states of attention. The related work will serve us as inspiration to formulate our research question exposed in Chapter 4.

In the fields of *visual deduction and detection,* Rudmann et al. (2003) record eye movements such as pupil size, position of the gaze, fixation duration and number of saccades, in order to detect in real time the cognitive status of participants based on the premise that participants are thinking about the object to which their eyes are directed. The task consisted of determining the direction of rotation for a target gear whose direction of turning is induced by the initial gear, as seen in Figure 3. Through the analysis of pupil size they estimate the emotional response and through the gaze position (on/off the screen) they detect distraction states. The findings from Rudmann et al. (2003) reveal the possibility to redirect the gaze strategically to improve HCI. However, they focus primarily on detecting states of attention rather than measuring cognitive load.

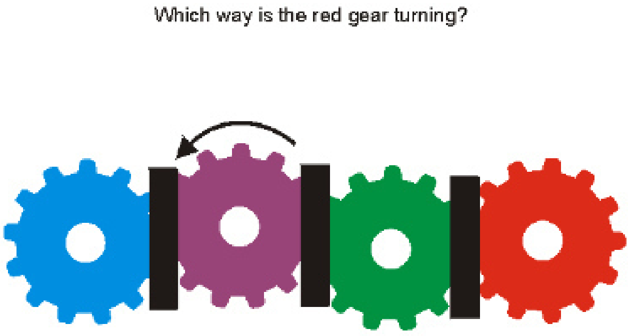


Figure 3: Discovering the rotation direction for a target gear in a simple cognitive task. [Rudmann et al., 2003]

In the same field, Pomplun and Sunkara (2003) investigated the relation between cognitive load and brightness by designing three tasks where red/blue squares and circles grew in size twice before they disappeared. Each task was performed at a different speed and participants had to click on the blue circle before it disappeared. Defining three levels of difficulty allowed them to study, thus, three levels of cognition. Their results are based on the pupillary response with two different levels of brightness. In order to compensate changes in the brightness and compute the pupil dilation induced only by cognitive load, they subtract the calibration value for the current display brightness of the current measured pupil size. However, they based their results relying only on the pupil dilation, leaving aside other eye tracking measurements that could have been of interest.

In the field of *memorizing*, Chen et al. (2011) design a task whose objective is to recall the positions of the defenders in a basketball game. In order to assess the level of cognition, they make use of the pupil size. Blinking rate and saccade velocity are used as an indicator of the mental effort. Moreover, they claim that an increment in fixation duration indicates an increment in the attention. Their contribution to HCI will help to develop intelligent interfaces that take into account where the attention is directed and how much user's attention is occupied by the task.

Furthermore, Rafiqi et al. (2015) and Klingner et al. (2008) make use of the shape, magnitude, and duration of pupil dilation to describe cognitive load in tasks where participants have to memorize a sequence of digits and report it back. In their results, one can appreciate how the pupil dilates while participants are memorizing the sequence of digits, as their cognitive load increase, and how the pupil contracts as they report the digits back, in line with a decrement of their level of cognition.

Moreover, Klingner et al. (2008) conducted experiments in the fields of *arithmetics* and *auditory detection* as well. In the former one, participants had to type in the product of two numbers between five and nineteen. In the latter one, they had to listen to a counting from one to nineteen and notify when they found a mistake in the sequence. Their findings relate an increment in the pupil diameter attending to an increment in the cognitive state of the participant.

In the field of *e-learning,* García Barrios et al. (2004) designed their own framework, AdELE which goal is to detect tiredness and mental effort to support adaptive teaching and learning. They use blinking rate and pupillary response to detect whether the participant is tired or stressed and when he suffers a high cognitive load. Detecting such states, allows the framework to adapt the content accordingly in real-time.

Toyama et al. (2015) include the use of an eye tracker with new technologies such as an *augmented reality* system. They try to determine whether a user is engaged with virtual content in the virtual display or focused on the real environment by analyzing the cognitive state of the user. However, they describe passive cognitive states based only on saccades' frequency. It would be interesting to complement their results with other eye tracking measurements.

At last, Iqbal et al. (2004) try to assess accurately the level of mental workload in order to develop an *attention manager* able to detect the user's state of attention and the best notification time. They want to prove that pupil size correlates well with the mental workload not only for discrete, non-interactive tasks but also for interactive tasks such as reading comprehension, searching, mathematical reasoning and object manipulation. They validate the mental workload through a user's subjective rating and task completion time and correlate the mental workload with the pupil dilation. They calculate a baseline pupil size based on users' fixation on a blank screen for 10 seconds. However, from our point of view, this technique has its limitations. If the baseline pupil size is calculated under the brightness of a white screen and tasks are performed on a screen of different brightness, one cannot be sure that changes in the pupil size are produced only due to the mental workload and not by the influence of changes in the brightness of the screen.

In conclusion, much has been investigated in the fields of eye tracking and cognitive load that has risen our curiosity and interest to get a deeper understanding of the topic. From our state-of-the-art research, we notice a lack of exploration in fields such as visual search or dual-task methods, a reason that has motivated our research question described in the following chapter.

4. Research Question

Through the study of the state-of-the-art about eye tracking and cognitive load, we have gained knowledge about how cognitive load can be accessed using an eye tracker and how to analyze the gaze events.

Most researchers focus their studies on measuring cognitive load in fields such as e-learning, detection of patterns in auditory and visual tasks, arithmetic operations and memorization, etc. However, few has been investigated in topics such as visual search or dual-tasks. This lack of extensive exploration has raised our interest in these fields. In everyday life, we are always visually searching. If we are able to find the relation between our gaze movements and the level of cognition when searching, we can design cognitively efficient visual interfaces which are one of the goals in HCI. In this regard, the possibility to investigate more in depth how eye tracking can be applied to measure cognitive load in visual search was one of our main motivations to formulate our research question.

The second main motivation explores which measurements have been generally used to study cognitive load and states of attention. Most of the studies relate states of attention, such as tiredness or distractions with gaze positions (on/off the stimulus) by measuring single parameters such as blinks. Regarding cognitive load, the relation between pupillary response and level of cognition is widely extended. However, little is known about how the *combination* of those eye tracking measurements can help to gain a deeper understanding about how the level of cognition can be detected. In this regard, the possibility to study more than one single eye tracking measurement has served me as a motivation to further investigate the relation between eye tracking and cognitive load.

Thus, with the knowledge gained through our state-of-the-art research about the fields where eye tracking can be applied to and how cognitive load can be assessed, we formulate the following research question:

"How to measure cognitive load using eye tracking in visual search tasks?"

In order to answer this question, we collect and analyze different eye tracking measurements, such as *fixations*, *saccades*, *pupil dilation* and *blinks*. Additionally, we combine these metrics with a NASA TLX (Task-Load-Index) questionnaire (Hart and Staveland (1988)), allowing us to gather participants' subjective information about their task load level.

Through the collection of the eye tracking measurements and NASA TLX questionnaire, we intend to answer the following sub-questions:

- *How does the analysis of fixations help us to understand cognitive load?*
- *How does the analysis of saccades help us to understand cognitive load?*
- *How does the analysis of pupil dilation help us to understand cognitive load?*
- *How does the analysis of blinks help us to understand cognitive load?*

- *How well correlate the analysis of fixations, saccades, pupil dilation and blinks to the subjective NASA TLX questionnaire?*

The aforementioned measurements will be collected from the experiment using a remote eye tracking and a NASA TLX questionnaire in paper form. The **analysis** of those measurements will be used to answer our research question *"How to measure cognitive load using eye tracking in visual search tasks?"* Additionally, the **outcome** of the analysis will help us to answer the five sub-questions previously mentioned.

In the following section, the outcomes and recommendations of our pre-study will be presented.

5. Pre-study

This chapter summarizes the pre-study we have conducted, based on the research of state-of-the-art conducted in our seminar. It offers a brief description of the design, conduction and analysis of a pre-attentive search task. Furthermore, we suggest some recommendations to face the challenges and problems that we have encountered throughout the whole process.

5.1 Task

Our pre-study consists of the design, development, conduction and analysis of a pre-attentive visual search task (parallel process). In such a task, participants need to find a target item in a pool of other distracting items, differentiated by a maximum of one property such as colour, shape, size, orientation, etc. Our task should be simple enough to allow us to measure cognitive load using eye tracking in an exploratory way. By doing so, we can use it as a baseline to study in the future more complex search strategies.

As independent variables, we selected the colour and shape of the elements, creating three different tasks where the participant has to find the target by its *colour*, *shape* or by the combination of both. A representation of our target and distractors for each condition can be seen in Figure 4.

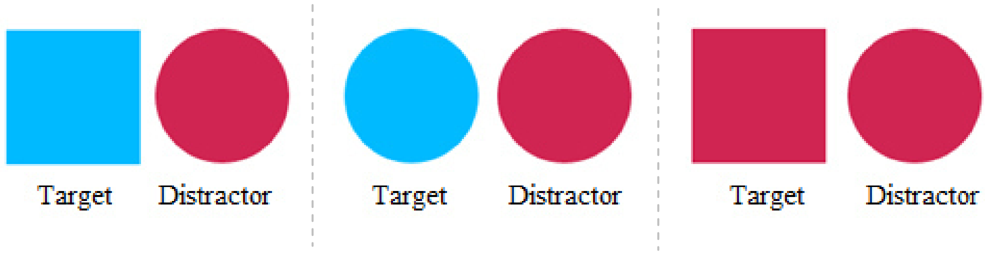


Figure 4: Levels of independent variable from left to right: Colour&Shape, Colour, Shape

The task is composed of thirty repetitions, divided in runs and lap. There are a total number of 10 runs per task and for each run, the number of distractors increases at 10% based on a linear increment of difficulty. Each run is divided, in turn, into 3 laps, where the number of distractors remains constant, in order to gather more data for the analysis. With thirty repetitions of the task, one can assure that the data collected represents each participant sufficiently.

Eye tracking measurements such as *fixations*, *saccades*, *blinks* and *pupil dilation*, are selected as our dependent variables. Their combined analysis provides a deeper understanding of the relation between cognitive load and eye tracking.

5.2 Setting

We set up the experiment in a controlled lab environment. The smart recorder was placed on a mobile desk of 1,05m high that was fixed at a distance of 1,30m away from the display. The monitor utilized was a Microsoft Perceptive Pixel with a 55 inch display and a resolution of 1920x1080p. In addition, a mouse to interact with the display was also placed on the desk. The position of the apparatus was kept constant across participants to assure that all of them performed the experiment under the same circumstances.

Equally important was to control the luminance of the environment, to assure that the reaction of the pupil was produced due to changes in the cognitive load and not due to changes in the luminance of the room. In order to control the luminance, all blinds in the lab were closed and the lights were turned on.

However, there was one parameter of the luminance that was not controlled and could imply changes in the pupil size, the luminance of the monitor. After the participant found the target element, the screen changed to a blank page, the so-called "resting screen" and whose objective was to redirect the gaze of the participant to the middle of the screen before a new lap started. With this screen, we induced a change in the luminance that could affect the results for the measurement pupil size.

5.3 Conduction

For the conduction of the experiment, we recruited ten participants, six male and four female. However, only data for four participants was utilized in the analysis.

When conducting an experiment with eye tracking, it is especially important to follow the same procedure with all participants, in order to assure that everyone performs the experiment under the same conditions, as eye tracking involves many steps. In this regard, it was especially useful to have a script to follow when explaining the experiment to the participants, to be sure that no step is forgotten. The procedure was carried out as follows:

1. The participants were welcomed and they were instructed to sign a declaration of consent and fill out a demographic questionnaire.
2. While wearing the eye tracker, the participants were introduced to the task.
3. Afterwards, the eye tracker was calibrated.
4. The participants performed the three tasks in a random order, to counterbalance. After the completion of each task, they filled out a NASA TLX questionnaire.
5. At last, they were asked to fill out a post-questionnaire and they were monetarily compensated for their time.

5.4 Analysis

The data recorded by the eye tracker were imported into the SMI BeGaze software, obtaining three different video and audio files, one per condition (colour, shape and colour&shape). These videos had to be analyzed frame by frame to identify the beginning of each lap manually. This implied a huge amount of time invested and for larger groups of participants was not a viable solution. In our experiment, we now log that information automatically, in a way that does not require analyzing the video to obtain that information.

Afterwards, the data was imported into Microsoft Excel, where it had to be manually formatted in order to operate with it. This issue obligated to invest a great amount of time and, as for the data export, the manual formatting was not a viable solution. In order to solve this issue, we have created scripts to automate this process in our current experiment.

Data for six participants was not considered appropriate for the analysis due to, for example, noisy data or failure in the recording process. Performing a statistical analysis over only four participants was, therefore, not viable.

At last, as each participant required a different amount of time to perform the task; its duration had to be normalized in order to make the tasks comparable across participants. The difference in task duration had a second consequence: the number of gaze events differed in each lap and for each participant. In order to average the data across participants and condition, it had to be normalized and the missing values had to be interpolated. To overcome this issue, we have selected a fixed lap duration in our new experiment.

5.5 Lessons learned

Through the design and conduction of our pre-study, we have faced many challenges that have helped us learn how to improve the design, conduction and analysis in further experiments.

One can summarize the lessons learned into four major problems. The first challenge is the selection of the task, a pre-attentive task, which was not demanding enough. Therefore, for our experiment we have decided to implement a conjunction task, where distractors differ among themselves, increasing the extraneous cognitive load.

The second main challenge is the selection of dependent and independent variables. For the dependent variables, one should assure that they are representative enough and in the case of blinks they were not, due to the short duration of the task. In most evaluation-based projects, independent variables are defined and fixed by the system to study, for example, screen size or input modality. However, to define independent variables in cognitive load is especially difficult as one has to design two tasks that should differ in the level of cognition. Therefore, one has to think carefully about the tasks to assure that the levels of independent variables represent different levels of cognition.

Studying the pupil dilation implies having a controlled lab environment. This is the third main challenge to face. One not only has to control the luminance of the environment, but also the luminance produced by the devices used. The luminance of the environment had been controlled by closing the blinds and turning on the light, but the luminance of the display was not taken into account. In our experiment, we redesign our tasks in order to keep the brightness of the display constant.

A good design and conduction of an experiment are worthless if one is not able to analyze the data in a proper and efficient way. The last challenge that we had to face was a too complex analysis and lack of participants to perform a statistical analysis. In our new experiment, we have partially automated the analysis. In this regard, we have included substantial information in a log file, such as lap's start/end timestamps and timestamps when the participant clicks on the target item. Furthermore, we have combined the data files and the log file into a single file to speed up the analysis and we have created scripts to automate the extraction and analysis of the data.

Furthermore, the experiment is designed to be performed with a minimum of thirty participants, in order to assure enough eye tracking data and to compensate for corrupted or too noisy data that should be left out of the analysis.

With the lessons learned and new ideas to face the challenges imposed by eye tracking and cognitive load, we redesign our experiment. The new proposal is described in the following chapters.

6. Experimental Design

This chapter summarizes the study description and setup of the experiment. Based on the literature review and related work, on the knowledge acquired through our pre-study and on our previous results, we redefine our procedure and the selection of tasks to study the relation between cognitive load and eye tracking.

6.1 Task description

Our goal is to design visual search tasks to answer our research question:

"How to measure cognitive load using eye tracking in visual search tasks?"

The objective of the tasks is to find a specific element (target item) among some other different elements (a pool of distractors) relying on the differences between target and distractors based on three different conditions: allocating the target by its colour – condition *colour* -, distinguishing it by its shape – condition shape – or identifying it by the combination of colour and shape – condition *colour&shape*. These three conditions are our independent variables.

The motivation behind choosing three conditions is to generate meaningful data about the differences between the gaze events that are my dependent variables. Additionally, having more than two independent variables will allow us to study the differences between the three conditions.

From our pre-study, we conclude that pre-attentive tasks designed were not demanding enough to withdraw clear conclusions about the relation between cognitive load and our dependent variables (fixations, saccades, blinks and pupil dilation). In this regard, we have modified the task developing a more demanding one. Instead of a pre-attentive task, where all distractors differ only by one condition from the target element, we implemented a conjunction task, where distractors differ from the target element but also they differ from each other.

For example, in condition *colour,* participants will have to identify a blue item that could take the shape of a square, circle or triangle, but with a **blue colour** among distractors that can be squares, circles or triangles in different colours but blue.

Likewise, in condition *shape* the objective is to find a target item with a **circle shape** in any colour among distractors that can be squares or triangles in any colour.

In condition *colour&shape* we create a combination of the other two conditions. In this case, participants have to find a target item with a **blue colour** and a **circle shape** among distractors that can be squares and triangles in any colour, including blue, or circles in any colour but blue.

A diagram with the possibilities can be seen in Figure 5.

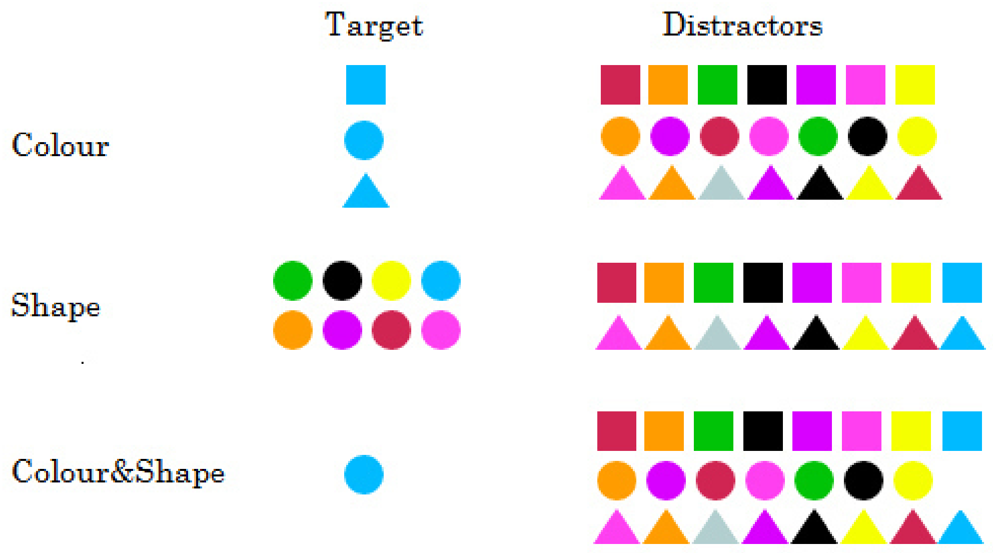


Figure 5. Description of target and distractors possibilities for each of the three conditions, colour, shape and colour&shape.

Our dependent variables are chosen to answer what we intend to study: if and how to measure cognitive load using an eye tracker through gaze events:

- Fixations.
- Saccades.
- Blinks.
- Pupil dilation.

As our literature review described in Chapter 3 suggests, there exists a relation between these gaze events and the level of cognition. Additionally, we want to study how well the results obtained in the analysis of gaze events match with what participants think about their level of cognition, in an indirect and subjective way. In this regard, we include a fifth dependent variable, the NASA TLX questionnaires, intended to measure task load. The aim is to find out if the subjective results from NASA

TLX are in line with the results obtained for each of our dependent variables and for each condition. NASA TLX derives an overall workload score based on ratings on six subscales, which are:

- **Mental demand:** How much mental and perceptual activity, such as thinking, searching, calculating, remembering, etc. does the participant need?

- **Physical demand:** How much physical activity in terms of moving, pushing, pulling, turning, etc., does the participant require?

- **Temporal demand:** How much time pressure does the participant feel due to the pace of the activities?

- **Performance:** How successful is the participant in accomplishing the goals of the task?

- **Effort:** How hard does the participant need to work (physically and mentally) in order to accomplish the task?

- **Frustration:** How stressed/relaxed, insecure/secure, irritated/content or discourage/gratified does the participant feel during the performance of the task?

In order to study if and how cognitive load changes when increasing the set size, we design the tasks to linearly increase the number of distractors towards the end of the task, for each of the conditions.

A summary of the workflow of the session to perform the tasks would be as follows:

1. The participant is situated wearing an eye tracker to gather information about our dependent variables (fixations, saccades, blinks and pupil dilation).

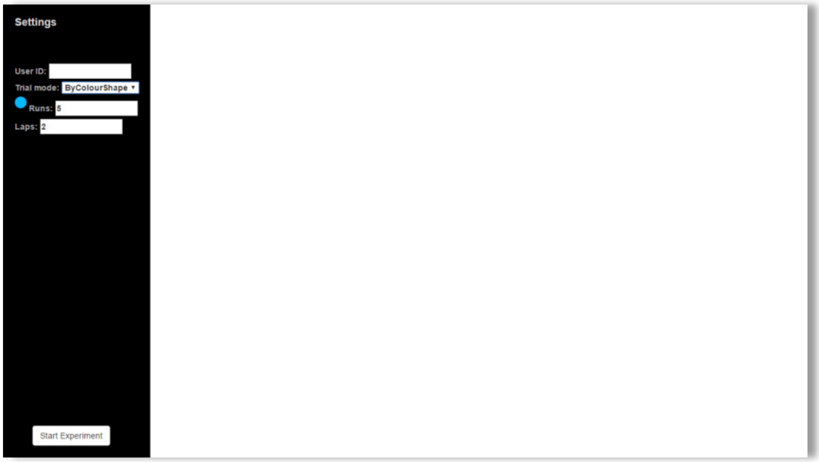
2. The participant performs three tasks consisting of finding a target element among a pool of distractors in several iterations where the number of distractors is linearly increased. Each of the tasks corresponds to our independent variables (conditions *colour*, *shape* and *colour&shape*).

3. Right after each task, the participant fills out a NASA TLX questionnaire to assess, in a subjective way, their task load, adding an extra value to our dependent variables.

In the following section, the implementation of these visual search tasks is described in more detail.

6.2 Implementation

In order to implement the visual search task, a web technology has been chosen, offering versatility and easy development. Javascript has been utilized as a programming language, allowing an interactive and dynamic web. The interface has been designed using HTML5 and CSS3. The task is composed of two screens, the main screen in which the settings of the task are configurable and a task screen in which the task is presented to the participant.



Settings

User ID:
Trial mode: ByColourShape
Runs: 5
Laps: 2

Figure 6. The main interface of the task with a movable panel to adjust the parameters of the task.

In the main interface, Figure 6, the examiner can configure the id of the participant, the type of task (condition *colour*, *shape* or *colour&shape*) getting a representative figure – for condition *colour* a blue item, for condition *shape* a circle and for condition *colour&shape* a blue circle – to help the participant in his search for the right item. Additionally, the number of runs and the number of laps can be configurable. Once the participant is ready to begin the task, he can click on the button "Start Experiment" situated in the bottom-left corner of the movable panel, after which the task interface will show up.

The task interface, Figure 7, presents a number of distractors and a target element that the participant has to find. The number of distractors increases linearly every run. Once the target is found and the participant has clicked on it, an auditory signal indicates that the right item was found.

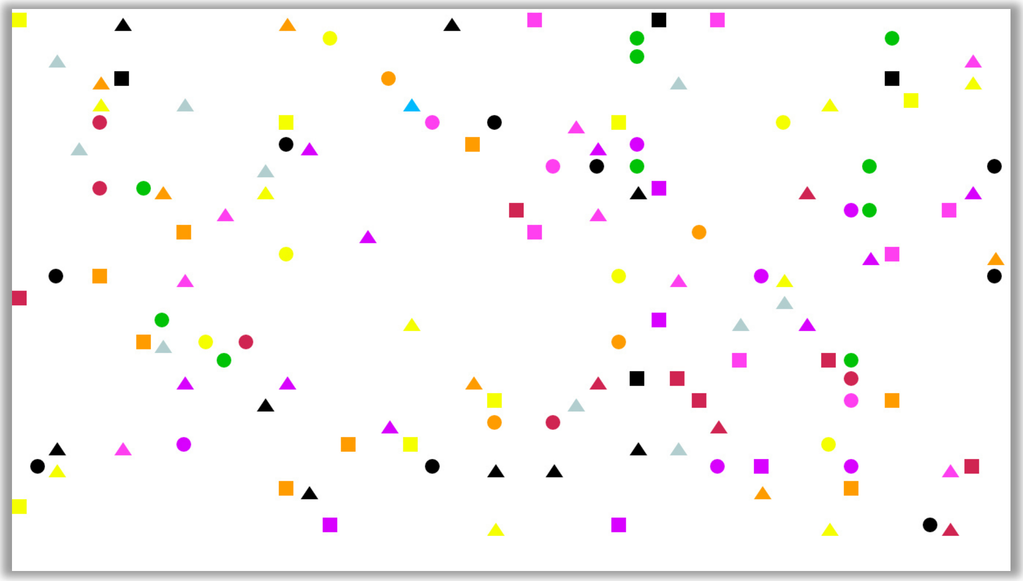


Figure 7. Screenshot for the condition Colour, where the participant has to find a blue element with a 10% of distractors presented.

Being the resolution of the display 1920x1080px and being the item size set to 20px, a grid to place the items in 54 columns and 29 rows is created. The number of items is calculated with the following Equation 1:

$$N^o\, of\ items = colums * rows = 54 * 29 = 1566 \qquad (1)$$

In order to increase the number of elements linearly, the number of runs is taken into account. Setting the number of Runs to 10, the number of elements will increase by 10% per run until covering the whole screen, attending to Equation 2:

$$Items\ increment = \frac{N^o\, of\ items}{runs} = \frac{1566}{10} \cong 156 \qquad (2)$$

An exception to this equation is the first run, for which one has to incorporate the target element. Therefore, the number of elements for Run 1 will be equal to the target item plus 155 distractors (one distractor less than "Items increment" to compensate the extra target element).

In an eye-tracking experiment, it is important to gather enough gaze events to make each run representative enough. Therefore, the variable "laps" is included. Setting Laps to 3, each run will be performed three times, keeping the number of items constant.

In order to find the right item, participants are provided with ten seconds, after which it is not possible to click on an element anymore and the screen changes to a resting interface. The objective of this resting screen is to redirect the gaze of the participant to the middle of the screen, assuring that all participants' gaze starting at the same position each lap. Additionally, for each run's resting screen, the number of runs left is indicated inside the circle. The resting screen is shown for two seconds, after which the next round with a new arrangement of the elements is presented. Furthermore, in this resting period, participants are instructed to relax and not focus on anything in particular. The objective is to detect changes in the level of cognition.

However, in eye tracking studies and especially when measuring the pupil dilation, keeping the luminance constant is crucial to obtain accurate results. In our pre-study, the resting screen presented a white interface with a circle in the middle of it, where participants had to direct their gaze. Changes in the luminance induced by changes in the brightness of the screen (from a colourful screen to a white one) could alter pupil size values measured by the eye tracker, which was altering the interpretation of the results. To deal with this obstacle, we introduced a resting screen where the luminance is kept constant by overlying the circle to redirect the gaze to the center of the screen (focus point) where the items are presented. This approach can be seen in Figure 8.

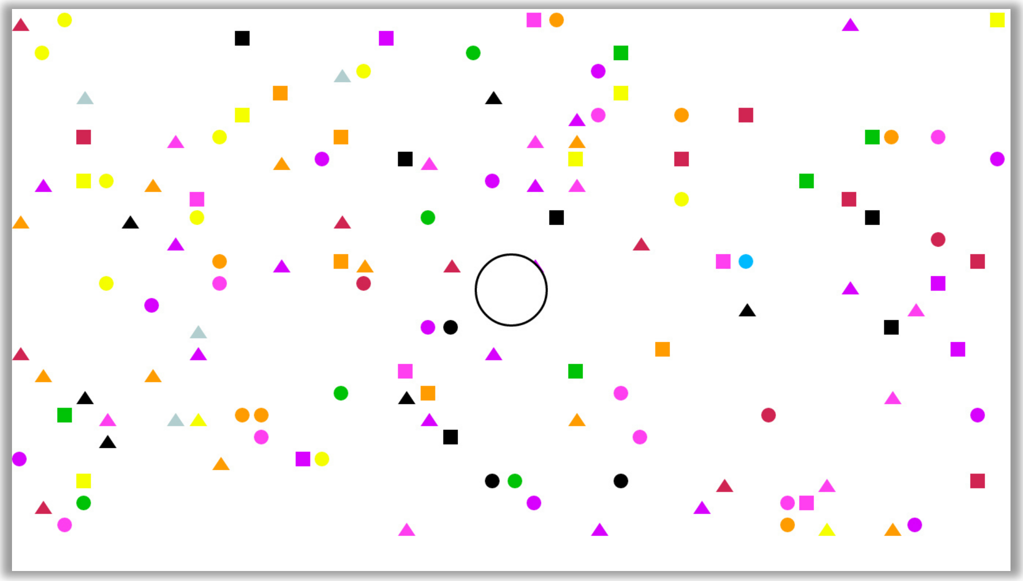


Figure 8. Resting screen to keep the luminance constant by overlying the focus circle to the previous task screen.

At last, a logging system is implemented, whose objective is to incorporate additional information to facilitate the subsequent analysis in a low-cost manner. The log file describes in detail insights of the iteration and helps to find traces of the task in relation to the gathered data from the eye tracker. This log file is created at the beginning of each task, collecting the following information:

- Participant ID.
- Condition: indicates which condition is being recorded, *colour*, *shape*, or *colour&shape*.
- Number of current lap.
- Number of current run.
- Start-lap: indicates the exact time where a lap started, in the format HH:MM:SS:MS

- End-lap: indicates the exact time where a lap ended, in the format HH:MM:SS:MS
- Click: indicates the exact time when the participant found the target item and clicked on it, in the format HH:MM:SS:MS

6.3 Apparatus

The setup of our experiment is kept in a controlled lab environment, having the advantage of a self-control of the settings, such as luminance regulation, fixed placement of the apparatus, etc., and reducing external influences that could distract participants' attention from their task.

Following the setup of our pre-study, we placed our recorder system on a mobile desk fixed at 1.20 meters high and separate 1.30 meters from the display (see Figure 9). The arrangement of the desk was slightly modified from its position in the pre-study in order to adjust it to the demand of some participants. In the pre-study, most participants claimed that the height of the desk was not the optimal one, generating an uncomfortable position for taller participants. In this regard, the height of the desk was increased by 0.15 meters. Additionally, the task was controlled using a wireless mouse. A keyboard was also placed on the mobile desk, although it was only used by the examiner.

As a display, we have chosen a 55 inches Microsoft Perceptive Pixel with a resolution of 1920x1080p and a brightness level of 400 nits. Being the mean height of participants 1.72 meters (STD = 0.09 meters; min = 1.60 meters; max = 1.90 meters), and in order to assure that participants' field of view covers the whole screen, we established the height of the display at 1.60 meters high, measured from the floor to the top of the display (see Figure 9).

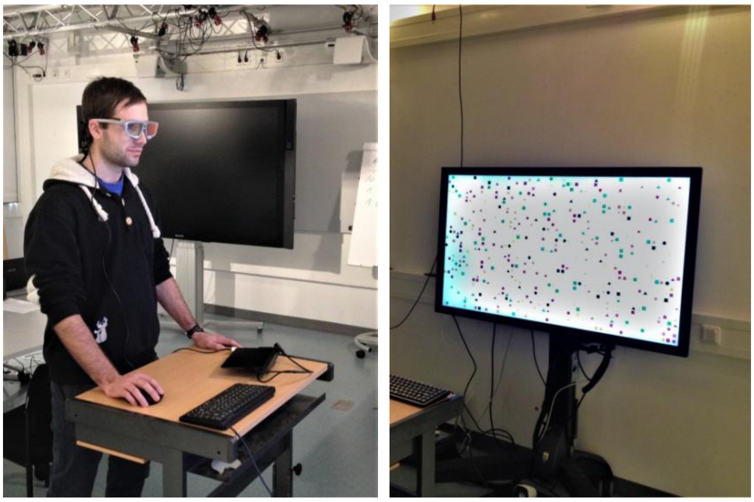


Figure 9. Setup of the mobile desk and eye tracker recorder system (left) and the Perceptive Pixel display (right)

Keeping the election of the recorder system the same as in the pre-study, we used a remote eye tracker, the SMI Eye-tracking Glasses 2, to record gaze events such as fixations, saccades, blinks and pupil dilation. According to Klingner et al. (2008), cognitive pupillometry can be extended from head-mounted systems to remote ones such as the SMI Glasses, deleting mobility restrictions present on head-mounted eye trackers. The eye tracker operates at 60HZ and provides a binocular gaze tracking over the participant's trackable field of view with an accuracy of 0.5° over all distances. The tracking range covers 80° horizontally and 60° vertically. Furthermore, it incorporates a scene camera with a resolution of 1280x960p at 24 fps or 960x720p at 30 fps. Its field of view is a bit more reduced than the eye tracker field, covering 60° horizontally and 46º vertically. Through the automatic parallax compensation, robust and accurate data are ensured over all distances.

The smart recorder is connected to the SMI Glasses and controlled remotely through the Experiment Centre software provided by SMI. The smart recorder has been upgraded from the pre-study, where a Samsung Galaxy Note 4 was used, to a Samsung Galaxy Tab. The main advantage lies in the possibility to connect the smart recorder via WiFi to a computer or tablet. This allows performing an accurate calibration process, it can be used to observe a live trace of participant's gaze and to add the participant's properties and annotations without intruding the participant's working area.

Although corrective lenses are provided we restricted their use, allowing only participants wearing contact lenses or no visually deficient. The reason behind it surfaces from our pre-study, where 40% of participants used corrective lenses and claimed that their performance was affected due to a mismatch between their current visual deficiencies and the corrective lenses. For example, some participants suffered from Astigmatism and Myopia and the use of corrective lenses could only compensate one of both deficiencies. In this regard, they could not see the screen properly affecting their performance. In order to assure that participants are not influenced by their visual deficiencies, we left out the use of corrective lenses.

Furthermore, in order to control the luminance of the room, avoiding external light influences that could derive in unexpected changes in pupil size, the blinds of the room were closed and the lights of the room were turned on.

Having a physical setup that is kept constant between participants' sessions allows us generating a comparable dataset for all participants. The complete setup of the experiment can be seen in Figure 10.

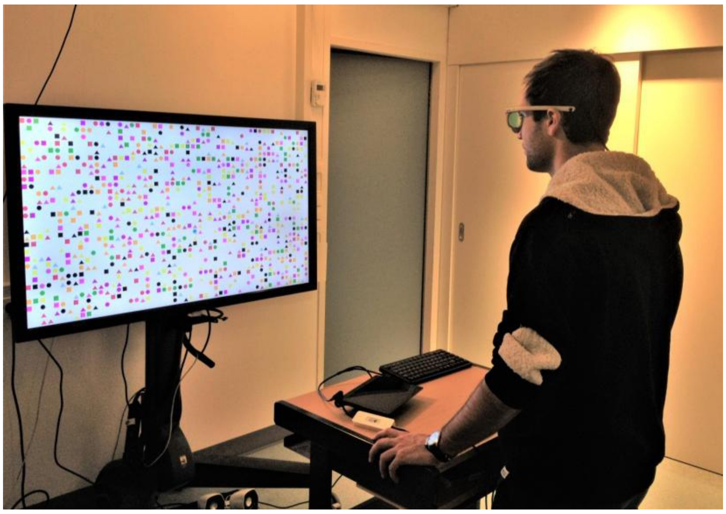


Figure 10. Setup of the experiment. The participant is wearing SMI Eye Tracker Glasses and he is situated on a mobile desk in front of the Microsoft Perceptive Pixel display.

6.4 Participants

Thirty-four participants were recruited to take part in the experiment; thirteen male and twenty-one female. The mean age was 24.03 years old (STD = 2.52 years old; min = 20 years old; max = 29 years old). Participants were college students at the University of Konstanz and with a level of education between bachelor and Ph.D. students. Students were from diverse fields of study, such as Economics, Psychology, Politics, Mathematics, Informatics, Teachers, Biology and Linguistics. It is worth mentioning that the different background of the participants reassures a diverted group of participants, which increases the representativeness of this study.

From our pre-study, we learned that it is very important to plan the experiment with a higher number of participants than the ones that are intended to account, to compensate for erroneous data that cannot be analyzed; or missing data due to failures in the recording process with the eye tracker. Consequently, we recruited 34 participants, counting on obtaining reliable data for, at least, 25 participants. The analysis of our data presented in Chapter 7 and the results reported in Chapter 8 were performed taking data from 26 participants after filtering invalid data.

Moreover, twelve participants usually wear glasses and for the experiment they made use of their own corrective contact lenses that are compatible with the SMI Eye Tracking Glasses, allowing them to use the eye tracker without adding the corrective lenses for it, but still not noticing deficiencies in their vision. Furthermore, none of the participants were colour blind and fourteen of them had previous experience using an eye tracker.

Participants were assigned to different conditions following a within-subjects design. This approach consists of assigning each participant to our three levels of the independent variable, say, each of the conditions (colour, shape and colour&shape). Although a between-subject design is easier to describe, smoother for participants and simpler to analyze, we decided to follow a within-subject design mainly due to, the number of participants. Working with data from twenty-six participants would not be significant enough if each participant performs only one condition.

Additionally, our goal is to detect possible variations in the level of cognitive load across conditions and for that, it is required that the same participant performs each of the conditions, as conditions for different participants would not be comparable. The main disadvantage of a within-subjects design is that each participant has to perform each of the tasks, lengthen the experiment time and being more tedious for them.

Another factor to take into account in a within-subject design is the so-called, learning effect. Participants performing tasks that are related can learn from them, affecting their further performance by gaining proficiency on them. One condition can influence in another condition if the task is very similar, as it is in our case. In order to deal with a learning effect issue, it is necessary to counterbalance conditions. Therefore, we made use of a 34 * 3 Latin square where, for each task and participant, one of the conditions – *colour*, *shape* or *colour&shape* – is assigned. Participants perform firstly Task 1, then Task 2 and at last Task 3. Each of the tasks corresponds to one condition, assigned in a random order, but making sure that every condition is performed the same number of times. A representation of the Latin-square can be seen in the following Table 1.

Table 1. Latin-square for the assignment of conditions to participants, where condition C1 = colour, C2 = shape and C3 = colour&shape.

	Task 1	Task 2	Task 3
Participant 1	C3	C1	C2
Participant 2	C1	C3	C2
Participant 3	C2	C3	C1
Participant 4	C3	C2	C1
Participant 5	C3	C1	C2
Participant 6	C2	C1	C3
Participant 7	C1	C3	C2
Participant 8	C2	C3	C1

Participant 9	C1	C2	C3
Participant 10	C3	C2	C1
Participant 11	C2	C1	C3
Participant 12	C1	C2	C3
Participant 13	C1	C2	C3
Participant 14	C1	C3	C2
Participant 15	C2	C1	C3
Participant 16	C2	C3	C1
Participant 17	C3	C2	C1
Participant 18	C3	C1	C2
Participant 19	C1	C2	C3
Participant 20	C1	C3	C2
Participant 21	C2	C1	C3
Participant 22	C2	C3	C1
Participant 23	C3	C2	C1
Participant 24	C3	C1	C2
Participant 25	C1	C2	C3
Participant 26	C1	C3	C2
Participant 27	C2	C1	C3
Participant 28	C2	C3	C1
Participant 29	C3	C2	C1
Participant 30	C3	C1	C2
Participant 31	C1	C2	C3
Participant 32	C1	C3	C2
Participant 33	C2	C1	C3
Participant 34	C3	C2	C1

TOTAL	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**
	12	11	11	11	12	11	11	11	12

After the experiment, participants received a fixed compensation of 8€ for one hour attendance to the study. The reason of fixing the compensation was to avoid that participants feel tempted to delay the end of the task, extending the session time and obtaining a higher compensation for it instead of performing their best. This awry motivation could lead to inaccurate data collection that would affect further results.

6.5 Procedure

It is widely known that having a formal procedure when conducting a study is highly beneficial. Especially when conducting eye tracking studies, it is important to instruct the participants systematically, assuring that no step is missed and that all participants receive the same information and guidance from the examiner. Hornbæk (2011) states that having a formal procedure helps reducing variability in the data, eases the collection of responses and the administration of treatments, and so forth. Therefore, we have prepared a structured procedure that is followed by each participant.

The examiner prepares the room, closing the blinds to assure constant luminance and prepares the eye tracker equipment. The eye tracking glasses are positioned on the mobile desk while the questionnaires are placed on a table. The examiner situates his equipment – a laptop connected via WiFi to the eye tracker –, behind the participant's mobile desk.

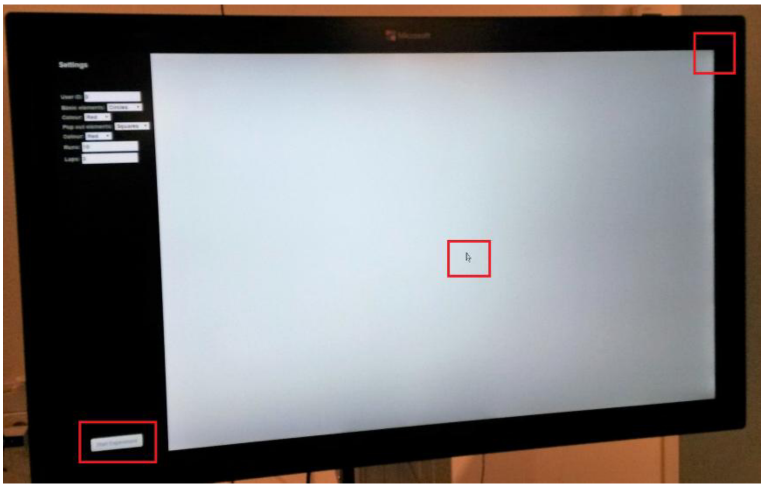
At the arrival of a participant, a welcome letter is handed out and the participant takes a seat on a table where he can read comfortably the letter. In addition, a declaration of consent regarding data collection and anonymous treatment the questionnaires, has to be signed. Participants are asked to fill a pre-questionnaire regarding demographic information, such as age, height, studies, possible vision deficiencies, as well as previous experience using an eye tracker. The welcome letter, declaration of consent and questionnaires can be found in Appendix A.

After filling the questionnaire, they are asked to move to the mobile desk, where they are introduced to the system, the SMI Tracking Glasses and the Perceptive Pixel display.

With the glasses on, the examiner proceeds to explain the upcoming tasks, what the target element is and how long they have to perform the task. It is especially important to remember the participants that they should redirect their gaze to the focus circle after each of the laps. By doing a short introduction, participants feel more comfortable and are encouraged to ask questions if they feel confused. Moreover, while the examiner exposes the goal of each task, the eye tracker is already gathering some data necessary to recognize and identify main points in the iris and pupil of the participant, allowing a calibration process.

Once the instruction of the task is completed and when the participant has no more questions, the examiner proceeds to calibrate the eye tracker. This means making sure that the gaze detection of the eye tracker reflects the exact same position in the real world. In this regard, we use a 3-point calibration that is controlled by the examiner remotely from his laptop. Participants are asked to look to three different landmarks and report it verbally when they are looking at them. The examiner will mark them with the help of his laptop, correcting any possible deviation detected in the gaze.

As landmarks, we have chosen the “Start experiment” button situated in the bottom-left corner of the screen, a mouse pointer situated previously in approximately the middle of the screen and the upper right corner of the screen. We choose these landmarks to assure that participants can identify them easily and to assure that the eye tracker can cover the totality of the display. The landmarks selected can be seen in Figure 11, highlighted with a red square.



Settings

Basic elements: Circles

Colour: Red

Pop out elements: Squares

Colour: Red

Figure 11. Highlighted landmarks in the bottom-left corner, middle of the screen and upper-right corner, used in a 3-point calibration.

Once the calibration is successful, the examiner instructs the participant to start the task whenever he feels ready. After performing each of the tasks, one per condition following a within-subject design, participants have to fill a NASA TLX (task-load-index) questionnaire and once they finish the last task they are asked to fill a post questionnaire about the use of the system (see Appendix A). It is especially useful to give them the chance to express any comments, rate the difficulty of the tasks in relation to one another (to compare them with the single-task rating from NASA TLX questionnaires) and to listen to any remarkable observations that they want to mention.

To conclude, participants are thanked for their time and are monetarily compensated for their participation. The conduction of the experiment lasts approximately 60 minutes.

7. Analysis

This chapter summarizes the methodology applied to analyze all data collected during our experiment. Firstly, we describe how to extract the eye tracking data using specific software for it and how to prepare such data in order to analyze it using statistical software. In the second section of this chapter, we describe different statistical tests applied to the data in order to test for differences between the conditions *colour*, *shape* and *colour&shape*.

7.1 Data preparation

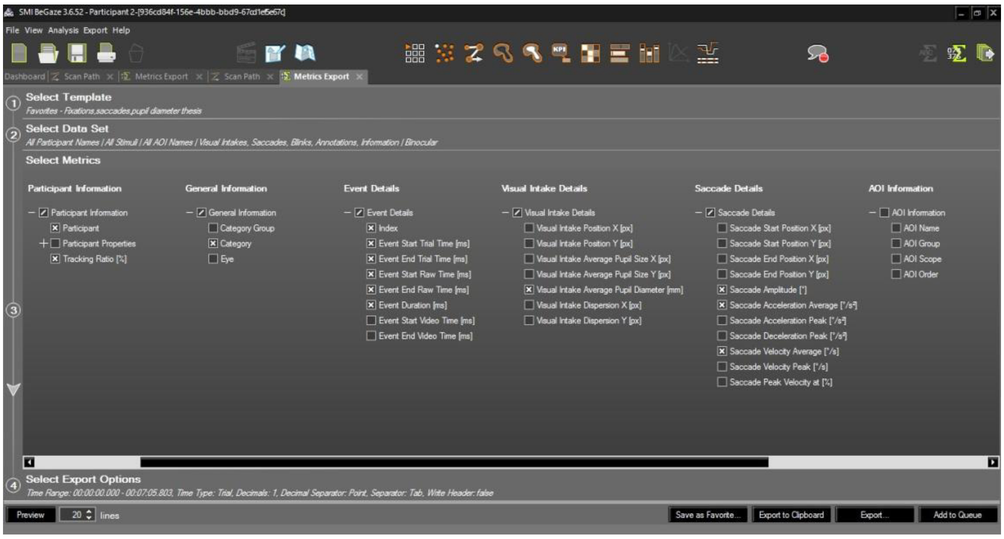
Once the data has been recorded using the eye tracker described in section 6.3 Apparatus, it has to be extracted. SMI provides a software package, the Behavioural and Gaze Analysis software (SMI BeGaze) that is fully integrated with the recording software utilized to collect the eye tracking data. The software assists with the analysis, visualization and export of gaze events.

For each participant three files are recorded by the eye tracker, corresponding to each of the conditions we want to study – *colour*, *shape* and *colour&shape* – which are imported into BeGaze software, obtaining three different audio and video files. BeGaze provides the possibility of replaying and examining the video of the experiment frame by frame and include annotations that can be helpful when analyzing the data. Furthermore, the software offers an Areas of Interest (AOI's) creator, visualizations of the data through gaze plots, such as scan paths and bee swarm plots, attention maps such as heat maps and focus maps. Additionally, it allows exporting the data in a RAW format or already processed gaze events to obtain them classified as Visual Intakes (fixations), Saccades and Blinks.

In our preparation of the data, it is necessary to replay the video for each of the conditions and manually identify the beginning of the experiment. In order to assure that no gaze events are missed due to a late start in the recording, we start it before the participant clicks on “Start Experiment”. This pre-data, therefore, has to be discarded and for that, we need to identify the real beginning of the task (clicking on “Start Experiment”) by replaying the video.

A relevant feature offered by BeGaze when replaying the video is the possibility to create, import and set timestamps, called annotations. These annotations are exported as a special type of gaze event, highlighting relevant points on the timeline, as the beginning of the task. For each of the conditions per participant, the video of the task was replayed and an annotation was added in the instant when the task starts. These annotations will help us to synchronize our eye tracking data file with our log file.

The Metrics Export option of BeGaze allows selecting per participant and per condition relevant information to be exported, such as Participant information, General Information, Event Details, Visual Intakes and Saccades details, and AOI information, as it can be seen in Figure 12.



1 Select Template
Favorites - Fixations,saccades,pupil diameter thesis

2 Select Data Set
All Participant Names | All Stimuli | All AOI Names | Visual Intakes, Saccades, Blinks, Annotations, Information | Binocular

3 Select Metrics

Participant Information

- ☑ Participant Information
  - ☒ Participant
  - ☐ Participant Properties
  - ☒ Tracking Ratio [%]

General Information

- ☑ General Information
  - ☐ Category Group
  - ☒ Category
  - ☐ Eye

Event Details

- ☑ Event Details
  - ☒ Index
  - ☒ Event Start Trial Time [ms]
  - ☒ Event End Trial Time [ms]
  - ☒ Event Start Raw Time [ms]
  - ☒ Event End Raw Time [ms]
  - ☒ Event Duration [ms]
  - ☐ Event Start Video Time [ms]
  - ☐ Event End Video Time [ms]

Visual Intake Details

- ☑ Visual Intake Details
  - ☐ Visual Intake Position X [px]
  - ☐ Visual Intake Position Y [px]
  - ☐ Visual Intake Average Pupil Size X [px]
  - ☐ Visual Intake Average Pupil Size Y [px]
  - ☒ Visual Intake Average Pupil Diameter [mm]
  - ☐ Visual Intake Dispersion X [px]
  - ☐ Visual Intake Dispersion Y [px]

Saccade Details

- ☑ Saccade Details
  - ☐ Saccade Start Position X [px]
  - ☐ Saccade Start Position Y [px]
  - ☐ Saccade End Position X [px]
  - ☐ Saccade End Position Y [px]
  - ☒ Saccade Amplitude [°]
  - ☒ Saccade Acceleration Average [°/s²]
  - ☐ Saccade Acceleration Peak [°/s²]
  - ☐ Saccade Deceleration Peak [°/s²]
  - ☒ Saccade Velocity Average [°/s]
  - ☐ Saccade Velocity Peak [°/s]
  - ☐ Saccade Peak Velocity at [%]

AOI Information

- ☐ AOI Information
  - ☐ AOI Name
  - ☐ AOI Group
  - ☐ AOI Scope
  - ☐ AOI Order

4 Select Export Options
Time Range: 00:00:00.000 - 00:07:05.803, Time Type: Trial, Decimals: 1, Decimal Separator: Point, Separator: Tab, Write Header: false

Preview 20 lines

Figure 12. BeGaze Metrics Export option with a selector of relevant information to be exported.

For each participant and each condition, we export a file containing metrics that are of relevance to assess cognitive load:

- **Participant**: Represents the name of the participant, expressed as P plus the id of the participant (for example, P25).
- **Tracking ratio**: An indication (in %) of how well the eye tracker detected and tracked participant's gaze.
- **Category**: Describes the category of the gaze event such as Visual Intake (fixations), Saccade, Blink or Annotation Instant.
- **Event Start Raw Time[ms]** and **Event End Raw Time[ms]:** Indicates the start/end times of the recording.
- **Visual Intake Average Pupil Diameter[mm]**
- **Saccade Velocity Average [°/s]**

A drawback of the software is the incapability to export the real time, not the relative time from Event Start/End Raw Time, expressed as the Date of Day - Time of Day from the already processed gaze events. We require this information in order to synchronize the gaze events with our log file. In this regard, it was necessary to obtain the Date of Day from the "Export Raw Data option" that correlates the Date of Day with the Event Start Raw Time. Three extra files were exported per participant, one per condition, to obtain this correlation.

As a matter of simplicity, when working with the eye tracking data it was necessary to merge both files to obtain the Date of Day information in the file that contains the eye tracking data. To do so, we use the software KNIME Analytics Platform that with its innovative node-system and more than 1000 modules eases the end-to-end analysis. Especially relevant to us are the modules intended for transforming the data. With a simple scheme, it is possible to merge tables of data by a common column, in our case "Event Start and End Raw Time" and integrate it afterwards by the "Date of Day" timestamps with our log file, incorporating relevant information such as Start/End Lap and Run time, the exact time when participant found and clicked on the target element and the number of the run/lap. By merging these three tables, we obtain a complete and self-content file with all the information we need for the subsequent analysis. The workflow can be seen in Figure 13.

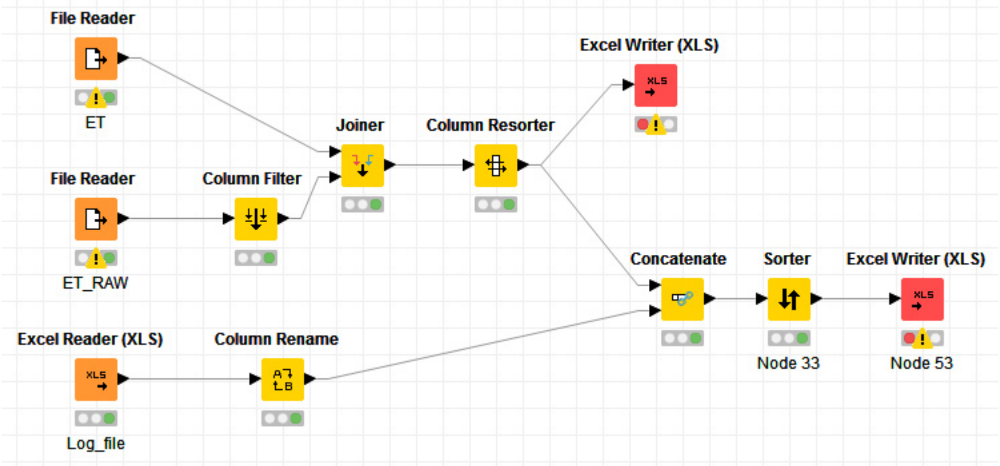
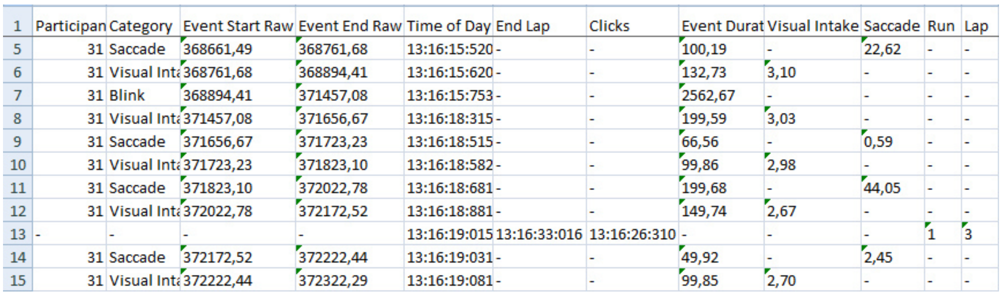


Figure 13. KNIME workflow for merging eye tracking exported files and log file.

Through the File reader and the Excel reader nodes, we import the data into the software. First, we filter our eye tracking raw data table to fetch the columns that are relevant for us ("Event Start and End Raw Time" and "Date of Day"), avoiding a heavier processing of the data in further operations. Next, we join both tables by their common column "Event Start Raw Table", in an inner join mode so only matching rows will show up in the output table. By doing this, we create a file with our gaze events processed and their corresponding "Date of Day" in the format HH:MM:SS,MS. This file is exported to an Excel format. At last, the joined table is concatenated with the log file by the column "Date of Day" and sorted in order to export it to an Excel file that contains all gaze events and the timestamps when each lap and run starts/ends plus timestamps when participants found the target item. An example of the Excel file can be seen in Figure 14.



1	Participan	Category	Event Start Raw	Event End Raw	Time of Day	End Lap	Clicks	Event Durat	Visual Intake	Saccade	Run	Lap
5	31	Saccade	368661,49	368761,68	13:16:15:520	-	-	100,19	-	22,62	-	-
6	31	Visual Int	368761,68	368894,41	13:16:15:620	-	-	132,73	3,10	-	-	-
7	31	Blink	368894,41	371457,08	13:16:15:753	-	-	2562,67	-	-	-	-
8	31	Visual Int	371457,08	371656,67	13:16:18:315	-	-	199,59	3,03	-	-	-
9	31	Saccade	371656,67	371723,23	13:16:18:515	-	-	66,56	-	0,59	-	-
10	31	Visual Int	371723,23	371823,10	13:16:18:582	-	-	99,86	2,98	-	-	-
11	31	Saccade	371823,10	372022,78	13:16:18:681	-	-	199,68	-	44,05	-	-
12	31	Visual Int	372022,78	372172,52	13:16:18:881	-	-	149,74	2,67	-	-	-
13	-	-	-	-	13:16:19:015	13:16:33:016	13:16:26:310	-	-	-	1	3
14	31	Saccade	372172,52	372222,44	13:16:19:031	-	-	49,92	-	2,45	-	-
15	31	Visual Int	372222,44	372322,29	13:16:19:081	-	-	99,85	2,70	-	-	-

Figure 14. Excel file containing all gaze events and timestamps.

In order to analyze the data using statistical software, such as SPSS, the data has to be formatted. For that, we use a script created using the MATLAB platform.

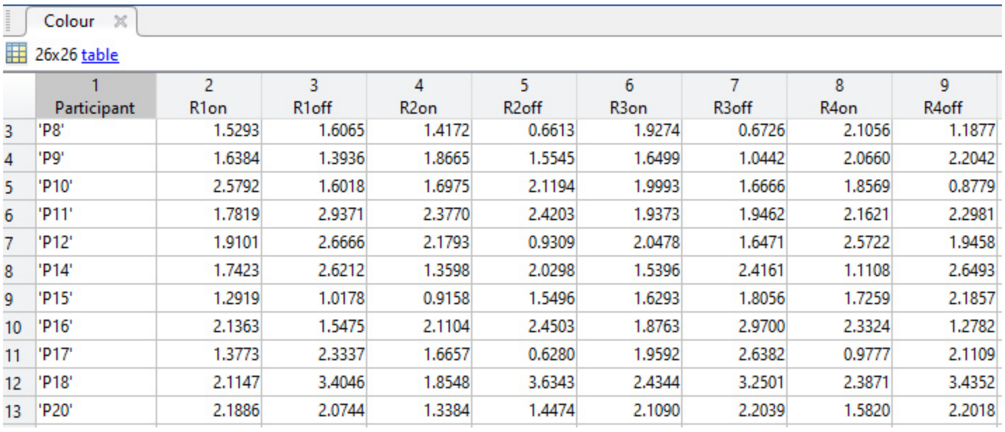
The main goal is to analyze the information at a Run level, making a distinction between gaze events obtained before the participant finds the target element (OnRun) and gaze events after the participant finds the element and before the next Lap starts (OffRun). Through the script, we extract all gaze events OnRun and OffRun for each of the ten runs, classifying them in Visual Intakes (fixations), Saccades and Blinks, and counting the number of gaze events present in each of the phases. Firstly, we need to convert our date format exported from BeGaze as HH:MM:SS,MS into a date format that is readable by Matlab. From our log file, we obtain the exact time when each lap started and ended, and when the element was found. With these three variables we can establish two time periods:

- **OnLap:** from the instant when lap started to the instant when the participant clicked on the target element.
- **OffLap:** from the instant when the participant clicked on the target element to end of the lap.

For each lap period, we obtain the number of gaze events. As the length of each period differs from participant to participant –

depending on how long they need to find the element – the data has been normalized calculating the number of gaze events per millisecond.

Each run is composed by three laps. In this regard, to calculate the number of gaze events OnRun and OffRun, we averaged the values of each three OnLap and OffLap values. Following this procedure, we calculate OnRun and OffRun values for the ten runs that the participant had performed on the task. An example of the structure for the output table can be seen in Figure 15.

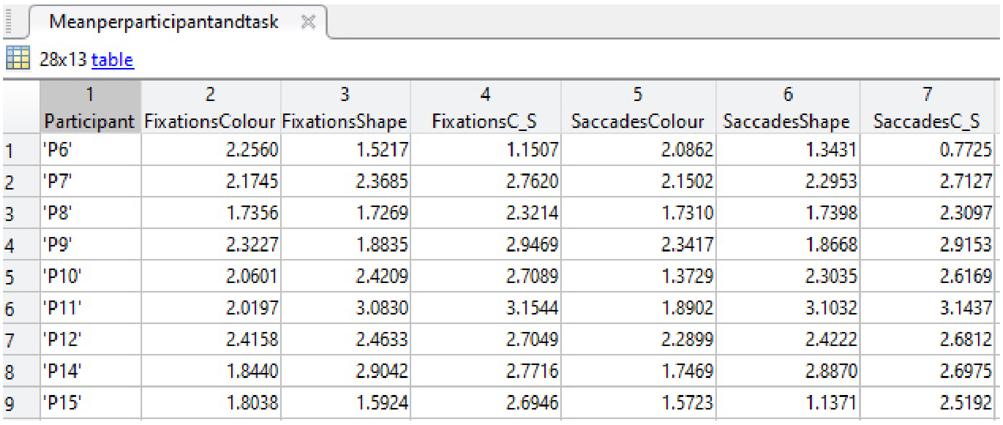


	1 Participant	2 R1on	3 R1off	4 R2on	5 R2off	6 R3on	7 R3off	8 R4on	9 R4off
3	'P8'	1.5293	1.6065	1.4172	0.6613	1.9274	0.6726	2.1056	1.1877
4	'P9'	1.6384	1.3936	1.8665	1.5545	1.6499	1.0442	2.0660	2.2042
5	'P10'	2.5792	1.6018	1.6975	2.1194	1.9993	1.6666	1.8569	0.8779
6	'P11'	1.7819	2.9371	2.3770	2.4203	1.9373	1.9462	2.1621	2.2981
7	'P12'	1.9101	2.6666	2.1793	0.9309	2.0478	1.6471	2.5722	1.9458
8	'P14'	1.7423	2.6212	1.3598	2.0298	1.5396	2.4161	1.1108	2.6493
9	'P15'	1.2919	1.0178	0.9158	1.5496	1.6293	1.8056	1.7259	2.1857
10	'P16'	2.1363	1.5475	2.1104	2.4503	1.8763	2.9700	2.3324	1.2782
11	'P17'	1.3773	2.3337	1.6657	0.6280	1.9592	2.6382	0.9777	2.1109
12	'P18'	2.1147	3.4046	1.8548	3.6343	2.4344	3.2501	2.3871	3.4352
13	'P20'	2.1886	2.0744	1.3384	1.4474	2.1090	2.2039	1.5820	2.2018

Figure 15. An example of the structure of an OnRun/OffRun table for the condition colour and the metric fixations. Columns of OnRun/OffRun five to ten are included in the table, although they are not presented in this example.

This process is done a total of nine times per participant (one file per condition and per gaze event). This file will help us to study whether there are significant differences between gaze events in periods where participants search for the item (presumably higher cognitive load), called OnRun periods, and in periods after the search (presumably low cognitive load) called OffRun, as well as to identify if there are significant differences between the beginning of the task and the end of the task, as the number of elements present on the screen increases linearly.

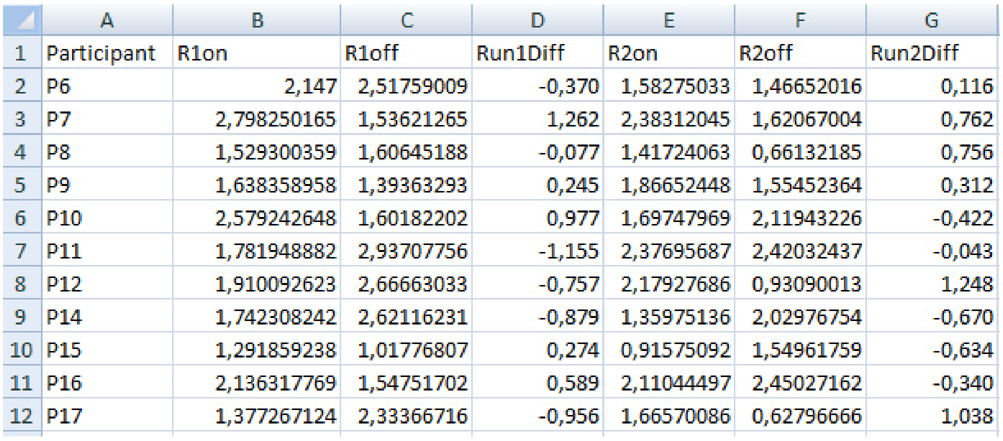
Additionally, we create an extra file with an average number of gaze events (fixations, saccades, blinks) per task and per condition, and the results of the NASA TLX questionnaires, in order to study if there are significant differences between our three independent variables, colour, shape and colour&shape. An example of the summary file can be seen in Figure 16.



	1	2	3	4	5	6	7
	Participant	FixationsColour	FixationsShape	FixationsC_S	SaccadesColour	SaccadesShape	SaccadesC_S
1	'P6'	2.2560	1.5217	1.1507	2.0862	1.3431	0.7725
2	'P7'	2.1745	2.3685	2.7620	2.1502	2.2953	2.7127
3	'P8'	1.7356	1.7269	2.3214	1.7310	1.7398	2.3097
4	'P9'	2.3227	1.8835	2.9469	2.3417	1.8668	2.9153
5	'P10'	2.0601	2.4209	2.7089	1.3729	2.3035	2.6169
6	'P11'	2.0197	3.0830	3.1544	1.8902	3.1032	3.1437
7	'P12'	2.4158	2.4633	2.7049	2.2899	2.4222	2.6812
8	'P14'	1.8440	2.9042	2.7716	1.7469	2.8870	2.6975
9	'P15'	1.8038	1.5924	2.6946	1.5723	1.1371	2.5192

Figure 16. An example of the structure of a summary table containing the averaged values for each gaze event per condition. The gaze event Blinks and results for NASA TLX are included in the summary table, although they are not presented in this example.

The statistical analysis is conducted using SPSS Statistics Base. The tests performed on the data are detailed in the following section. However, before conducting some of the tests such as Normality test or dependent T-Test, the OnRun/OffRun file has to be extended, adding new columns with the differences between the On values and Off values for each of the runs (RunDiff = OffRun – OnRun). This is done manually using excel and an example of the resulting file can be seen in Figure 17.



	A	B	C	D	E	F	G
1	Participant	R1on	R1off	Run1Diff	R2on	R2off	Run2Diff
2	P6	2,147	2,51759009	-0,370	1,58275033	1,46652016	0,116
3	P7	2,798250165	1,53621265	1,262	2,38312045	1,62067004	0,762
4	P8	1,529300359	1,60645188	-0,077	1,41724063	0,66132185	0,756
5	P9	1,638358958	1,39363293	0,245	1,86652448	1,55452364	0,312
6	P10	2,579242648	1,60182202	0,977	1,69747969	2,11943226	-0,422
7	P11	1,781948882	2,93707756	-1,155	2,37695687	2,42032437	-0,043
8	P12	1,910092623	2,66663033	-0,757	2,17927686	0,93090013	1,248
9	P14	1,742308242	2,62116231	-0,879	1,35975136	2,02976754	-0,670
10	P15	1,291859238	1,01776807	0,274	0,91575092	1,54961759	-0,634
11	P16	2,136317769	1,54751702	0,589	2,11044497	2,45027162	-0,340
12	P17	1,377267124	2,33366716	-0,956	1,66570086	0,62796666	1,038

Figure 17. An example of an extended OnRun/OffRun file prepared to be used with a statistical analysis software such as SPSS, for the condition colour and the metric fixations.

7.2 Data analysis

In order to study the relation between cognitive load and eye movements, we focus on answering the four research sub-questions:

- *How does the analysis of fixations help us to understand cognitive load?*
- *How does the analysis of saccades help us to understand cognitive load?*
- *How does the analysis of pupil dilation help us to understand cognitive load?*
- *How does the analysis of blinks help us to understand cognitive load?*

- *How well relate the analysis of fixations, saccades, pupil dilation and blinks to the subjective NASA TLX questionnaire?*

The statistical analysis of the gaze data collected with the eye tracker, together with the subjective task load measurements collected through the NASA TLX, are analyzed using statistical software from IBM, which is called SPSS Statistics Base.

For each of the eye tracking measures and for each of the conditions – *colour*, *shape*, *colour&shape* – we want to find out:

1. If there are significant differences between the measures while participants perform the task (OnRun) and the measures in the resting state, while participants wait for the next run to start (OffRun).
2. If there are significant differences between the averages in the gaze events across the conditions for the whole task, irrespective of the run.

We accomplish the former using a dependent T-Test to compare the means between our two related groups (RunOn-RunOff) on our dependent variable (fixations, saccades, blinks or pupil dilation) for each of the ten runs.

However, in order to analyze the data using a dependent T-Test, one has to make sure that the data can fulfill the requirements for a dependent T-Test (Field (2009)). This means that, in order to obtain a valid result, the data have to fulfill four assumptions:

- Assumption 1: The **dependent variable** should be measured on a **continuous scale**. That means it has to be a quantitative variable. Our gaze events are measured in number of events/millisecond, fulfilling the first assumption.

- Assumption 2: The **independent variable** should consist of two **related groups**, being the same number of subjects present in both groups. This means that each participant has been measured on two occasions on the same dependent variable. In our case, our independent variable (number of Run) consists of two related groups, the first related group measuring the gaze events while performing the task (OnRun) and the second related group measuring the gaze events after finding the element and while waiting for the next Run to start, in the resting period (OffRun). In addition, the same participants have been measured on both occasions, while performing the task (OnRun) and in the resting period (OffRun).

- Assumption 3: There should be **no significant outliers** (data points that do not follow the usual pattern) on the **differences** between the two related groups. This assumption is especially important to fulfill since the statistical significance of the test can be affected, reducing the validity of the results. In order to detect outliers, we calculate the difference between the related groups for our ten Runs:

$$DiffRun_i = RunOff_i - RunOn_i \qquad i \in \{1..10\} \qquad (3)$$

Using SPSS we obtain a Histogram and a Normal Q-Q Plot from the difference between the related groups, which will help to identify outliers. In Figure 18 one can observe a possible outlier in the Histogram for a value of -2,0 and in Figure 19 one can distinguish a possible outlier in the Q-Q Plot, marked in red, that deviates from the distribution.

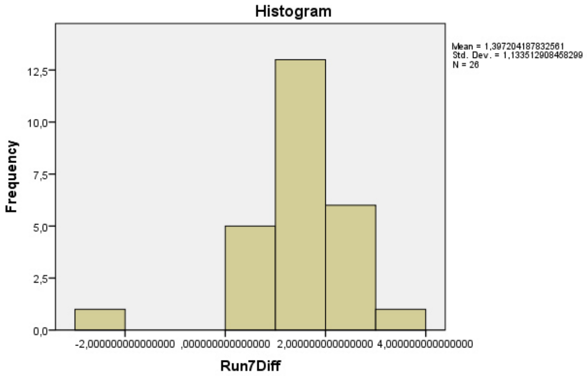


Figure 18. Histogram from the difference in the related groups for Run 7. One can observe a possible outlier for values under -2,0.

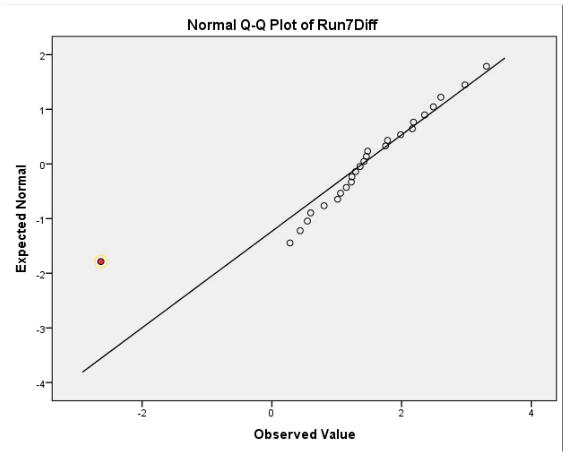


Figure 19. Q-Q Plot from the difference in the related groups for Run 7. One can observe a possible outlier that deviates from the distribution on the left side of the plot, marked in red.

Once we identify possible outliers, we use the 'outlier labeling rule', consisting on multiplying the Interquartile Range by a factor of 2.2 (Hoaglin and Iglewicz (1987)). The descriptive statistics of SPSS provide a Percentile table with the values we need. To calculate the upper and lower limits outside of which the data is identified as outlier, we use the following formulas:

$$UpperLimit = \mathrm{Q3} + (\mathrm{f} * (\mathrm{Q3} - \mathrm{Q1})) \tag{4}$$

$$LowerLimit = \mathrm{Q1} - (\mathrm{f} * (\mathrm{Q3} - \mathrm{Q1})) \tag{5}$$

being f our factor with a value of 2.2, Q1 the 25 Percentile and Q3 the 75 Percentile.
SPSS provides, as well, a list of the fifth highest and lowest Extreme Values, which should belong to the range set by UpperLimit and LowerLimit. In case any value breaks the limits, it should be excluded when performing the T-Test.
Our data has been checked to detect and eliminate the presence of outliers.

- Assumption 4: The **distribution of the differences** in the **dependent variable** between the two related groups – DiffRun – should be **approximately normally distributed**. Due to the robustness of the dependent T-Test, small violations of the assumption still provide valid results. However, to avoid defining "small violations" in an inaccurate manner that might lead to erroneous results; we perform the dependent T-Test only on data that is completely normally distributed.

In order to test for normality, we use a Shapiro-Wilk test. This test is appropriate for small sample sizes and therefore, it is adequate for our 26 sample size, although it can also be used in larger sets up to 2000 samples. Using this test we numerically assess normality.

In Chapter 8., we report the statistics in the following format: *t* (degrees of freedom) = *t*-value; p = significance value. Values with a significant level under 0.05 reject the hypothesis that the sample is from a normal population, observing significant differences between the values. Values with a significant level over 0.05 indicate that the data is normal. These values are marked with a green background in the summary of the statistical test table for each of the eye tracking measures.

One can test for normality graphically analyzing the output of a Normal Q-Q Plot. Data points close to the diagonal line are indicators that the data is normally distributed. On the contrary, if the data points stray from the line in an obvious non-linear fashion, then the data is not normally distributed. Figure 20 is an example of a Q-Q Plot where the data is normally distributed.

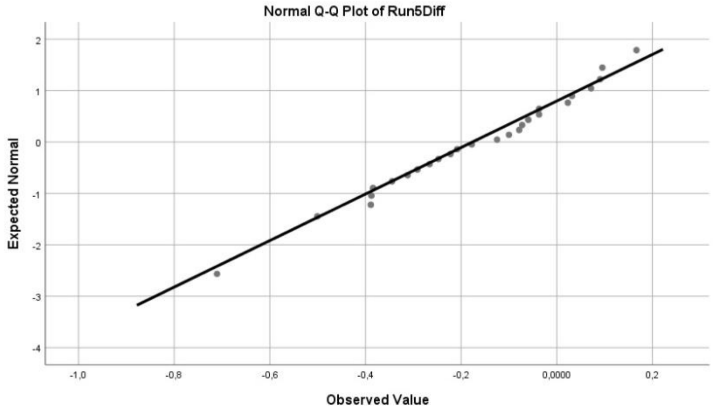


Figure 20. Normal Q-Q Plot for Run5 where data is normally distributed.

However, when working with eye tracking data, it is not uncommon that the fourth assumption of normality is violated. In these cases, a dependent T-Test should be substituted by its non-parametric equivalence, the Wilcoxon signed-rank test. In order to analyze the data using this test, it is also required to check that the data fulfill three assumptions (Field (2009)). The first two assumptions – continuous dependent variables and related independent variables – are common with the Shapiro-Wilk test and therefore, not explained in detail, as we have already seen that our data fulfills both of them. The third assumption, however, has to be checked in order to obtain valid results from the Wilcoxon signed-rank test:

- Assumption 3: The distribution of the **differences between the two related groups (DiffRun)** needs to be symmetrically shaped. In order to test this assumption, one can study the Boxplot, see Figure 21, to determine if the data is symmetrical in shape.

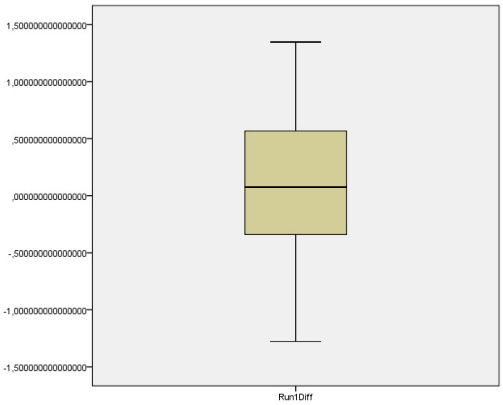


Figure 21. Boxplot from the distribution of the difference in the related groups for Run 1. One can observe asymmetry in the data, fulfilling the third assumption for a Wilcoxon signed-rank test.

Additionally, one can test for symmetry in the data by conducting a Frequency test and checking for skewness. SPSS defines an asymmetric distribution if “*a skewness value is more than twice its standard error*” and, therefore, the Wilcoxon signed-rank test should not be used. This symmetry can also be observed if we plot a Histogram with its normal curve, as in Figure 22.

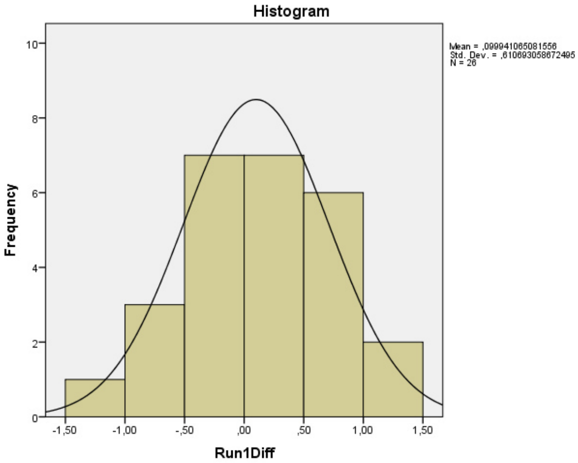


Figure 22. Histogram from the distribution of the difference in the related groups for Run 1. One can observe a normal distribution of the data, fulfilling the third assumption for a Wilcoxon signed-rank test.

A “repeated measures ANOVA” test is mainly used to compare changes in the response of participants to different conditions. In order to study if there are significant differences between the means in the gaze events across the conditions for the same participants, ANOVA with repeated measures is the right test. The results of the test are only valid if the data fulfills five assumptions

(Field (2009)), the first three are in common with a T-Test and therefore, just briefly mentioned.

- Assumption 1: The **dependent variable** should be measured on a **continuous scale**.

- Assumption 2: The **independent variable** should consist of two **related groups**, being the same number of subjects present in both groups.

- Assumption 3: There should be **no significant outliers** (data points that do not follow the usual pattern) on the **differences** between the two related groups.

- Assumption 4: The **dependent variable** in all groups to be compared should be **approximately normally distributed.** The data is tested for normality using a Shapiro-Wilk test. Although the repeated measures ANOVA is a robust test for small violations of normality, we only perform this test when the data is completely normally distributed.

- Assumption 5: The **variances of the differences** between all combinations of related groups (colour, shape and colour&shape) must be **equal**. This condition is known as **sphericity**. If this assumption is violated, it could lead to detecting significant differences where there are none. It is, therefore, especially important to test the data for sphericity through a Mauchly's test. We can assure that the assumption of sphericity has not been violated when there is no significant difference in the variances ($p > 0.05$).

The following section presents the results obtained through the conduction of the mentioned tests and a discussion about those results.

8. Results

This section presents the objective results obtained from each of the gaze events gathered using the eye tracker, as well as the subjective results obtained through each participant's NASA TLX questionnaires, for each of the three conditions – *colour*, *shape* and *colour&shape*. The goal is to identify if there are significant differences in the number of gaze events when the cognitive load increases, by increasing the number of distractors in each run. In addition, we want to study how a resting period after finding the items affects the participants' cognitive load. At last, we intend to find statistically significant differences between our three conditions – *colour*, *shape* and *colour&shape* – caused by changes in participants' cognitive load.

Based on the results, we will report for each of the measurements their applicability to measuring cognitive load.

8.1 Fixations

Fixations refer to the period of time where the eye is still, focused on a single point for a significant period of time. The variation on fixations has been reported (Rudmann et al. (2003), Chen et al. (2011)) to have a relation with attention levels and working memory. Particularly, they claim that factors such as fixation duration increases when an increment on the working memory is produced. But, how does the number of fixations reflect the level of cognition? Can fixations be used as an indicator to measure cognitive load?

In order to answer those questions, we study how the number of fixations per second changes from a period in which participants have to actively and visually search for an element target (higher level of cognition, OnRun) in comparison to a posterior period, where participants have already found the target element and have no further goals (lower level of cognition, OffRun). In this regard, we compare for each of the ten runs, if there is a statistically significant difference between the two mentioned periods, OnRun and OffRun.

Each of our conditions – *colour, shape* and *colour&shape* – is designed to present a different level of cognitive load. We divide our results about fixations attending to each of our conditions.

Colour

In order to compare On and Off runs, we make use of a dependent T-Test. To do so, the differences in the mean values of our OnRun and OffRun have to be normally distributed. Normality is tested through the Shapiro-Wilk Test. The following table, Table 2, present the results of these tests obtained for each of the runs, averaged by participants. Additionally, some descriptive statistics are presented. The highlighted data is formatted according to its relevance for us. A green background represents relevant data such as a normal distribution ($p>0.05$) in Shapiro-Wilk Test or significantly different data in our T-Test ($p<0.05$). A red background makes reference to data that is irrelevant for us. This scheme will be applied in all successive tables.

Table 2. Results for the gaze event *fixations* for condition *colour* averaged by participant, for each of OnRun/Off periods. Data is tested for normality and for statistically significant differences between those two periods of time.

N°Fix/s - Colour		**Descriptive Statistics**		**Shapiro-Wilk Test**	**Dependent T-Test**	
		Mean N°Fix/s	Std. Deviation	p-value	t-value	p-value
Pair 1	**R1on**	2,0129	0,390	0,825	-1,561	0,130
	R1off	2,2442	0,664			
Pair 2	**R2on**	1,8077	0,389	0,264	-1,434	0,164
	R2off	1,9878	0,693			
Pair 3	**R3on**	1,8543	0,238	0,787	-1,546	0,134
	R3off	2,0188	0,599			
Pair 4	**R4on**	1,9410	0,439	0,886	-0,252	0,803
	R4off	1,9792	0,671			
Pair 5	**R5on**	2,0265	0,337	0,696	-1,201	0,241
	R5off	2,1515	0,528			
Pair 6	**R6on**	2,0593	0,392	0,745	0,146	0,885
	R6off	2,0410	0,453			
Pair 7	**R7on**	2,1248	0,395	0,204	1,115	0,275
	R7off	1,9763	0,607			
Pair 8	**R8on**	2,1287	0,399	0,792	0,645	0,524
	R8off	2,0423	0,622			
Pair 9	**R9on**	2,1259	0,482	0,170	1,458	0,157
	R9off	1,8985	0,676			
Pair 10	**R10on**	2,2284	0,460	0,054	1,697	0,102
	R10off	1,9467	0,612			

Observing Table 2, we can study the distribution of the data. If the Significant value (p-value) of the Shapiro-Wilk Test is greater than 0.05, the data is normal. If it is below 0.05, the data significantly deviate from a normal distribution. For each of the ten pairs, we can observe that our p-values are greater than 0.05, what indicates that our data is normally distributed and it is adequate (green highlighting) to conduct a dependent T-Test.

Our dependent T-Test reveals that there are no significant differences (red highlighting) between the number of fixations OnRun and OffRun (p>0.05). Additionally, negative t-values indicates that the number of fixations per second is higher in OffRun periods while positive t-values indicate a higher number of fixations per second in OnRun periods.

Based on these results we can argue that from Run 6 on, positive values of our t-value (t-value>0) indicate an increment in the number of fixations OnRun, overtaking the number of fixations in OffRun periods, as participants need to focus on more items before finding the target one. However, for condition colour there are no significant differences between On/Off periods based on the number of fixations per second. In this regard, fixations cannot serve as an indication of the level of cognition, for a condition based on the colour of the target, where the visual search imposes fewer cognitive load.

Shape

For condition *shape* we have followed the same procedure as for condition *colour*. Likewise, we look for significant differences between active periods (OnRun) and resting periods (OffRun) and for that we make use of a dependent T-Test when the data is normally distributed or a Wilcoxon Test when it is not. The results obtained can be seen in Table 3.

Table 3. Results of the Shapiro-Wilk test dependent T-Test and Wilcoxon test for the gaze event *fixations* in condition *shape* averaged by participant, for each of OnRun/OffRun periods.

N°Fix/s - Shape		**Descriptive Statistics**		**Shapiro -Wilk Test**	**Dependent T-Test**		**Wilcoxon Test**	
		Mean N°Fix/s	Std. Deviation	p-value	t-value	p-value	Z-value	p-value
Pair 1	**R1on**	2,368	0,739	0,540	1,025	0,315		
	R1off	2,216	0,680					
Pair 2	**R2on**	2,063	0,680	0,234	1,532	0,138		
	R2off	1,907	0,658					
Pair 3	**R3on**	2,003	0,584	0,888	1,054	0,301		
	R3off	1,868	0,769					
Pair 4	**R4on**	2,250	0,700	0,013			-2,907	0,004
	R4off	1,735	0,794					
Pair 5	**R5on**	2,368	0,616	0,107	3,030	0,005		
	R5off	1,807	0,985					
Pair 6	**R6on**	2,255	0,732	0,206	3,256	0,003		
	R6off	1,739	0,829					
Pair 7	**R7on**	2,204	0,741	0,084	2,355	0,026		
	R7off	1,742	0,741					
Pair 8	**R8on**	2,253	0,704	0,453	4,420	0,000		
	R8off	1,574	0,757					
Pair 9	**R9on**	2,346	0,598	0,006			-3,484	0,000
	R9off	1,706	0,718					
Pair 10	**R10on**	2,163	0,532	0,678	2,795	0,010		
	R10off	1,787	0,619					

In Table 3, we can observe through the p-value of the Shapiro-Wilk Test that data for Run 4 and Run 9 are not normally distributed (p<0.05) while for the rest of the runs, the data follows a normal distribution (p>0.05). Normally distributed data is highlighted in green and it is suitable to conduct a dependent T-Test. Through our dependent T-Test and its non-parametric version, the Wilcoxon Test, we can conclude that the mean number of fixations per second differed statistically significantly between OnRun and OffRun periods for runs 4 onwards ($p<0.05$). This significant difference is highlighted in green. For earlier runs (Run1 to Run3) one cannot observe any statistically significant differences ($p>0.05$). This fact is indicated by a red background in the data in Table 3.

Furthermore, positive t-values in the dependent T-Test and negative z-values in Wilcoxon Test are an indicator of a higher number of fixations per second in OnRun periods, compared to the values in OffRun periods. This result is in line with the need to focus more repeatedly when searching for a target item than once the target has been found.

One can set a threshold indicating that from Run4 on the level of cognition can be related to the number of fixations per second in OnRun periods. If the number of distractors presented on screen (40% of distractors or more in this case) is high enough, it generates a level of cognition that can be detected through the difference in fixations per second in comparison to the resting period that follows once the participant has found the item.

In addition, we have found significant differences between the mean number of fixations per second for the condition *shape* but, however, we have not found any significant difference in condition *colour.* This fact suggests that condition *shape* induces a higher level of cognition than condition *colour*, which is in line with expectations when we designed both tasks. However, this fact will be investigated more in depth later in this section, including in the comparison the condition *colour&shape* described below.

Colour&Shape

For condition *colour&shape* we have followed the same procedure as for conditions *colour* and *shape*. Results obtained for normality test, dependent T-Test and Wilcoxon Test can be seen in Table 4.

Table 4. Results of the Shapiro-Wilk test, dependent T-Test and Wilcoxon test for the gaze event *fixations* and condition *colour&shape*, for each of OnRun/OffRun periods.

N°Fix/s Colour&Shape		**Descriptive Statistics**		**Shapiro-Wilk Test**	**Dependent T-Test**		**Wilcoxon Test**	
		Mean N°Fix/s	Std. Deviation	p-value	t-value	p-value	Z-value	p-value
Pair 1	**R1on**	2,194	0,649	1,000	0,834	0,412		
	R1off	2,094	0,741					
Pair 2	**R2on**	2,201	0,458	0,087	1,074	0,293		
	R2off	2,100	0,512					
Pair 3	**R3on**	2,587	0,640	0,896	5,068	0,000		
	R3off	1,762	0,516					
Pair 4	**R4on**	2,535	0,536	0,548	5,565	0,000		
	R4off	1,744	0,517					
Pair 5	**R5on**	2,718	0,549	0,993	7,568	0,000		
	R5off	1,281	0,700					
Pair 6	**R6on**	2,867	0,628	0,923	8,758	0,000		
	R6off	1,266	0,585					
Pair 7	**R7on**	2,787	0,571	0,004			-3,848	0,000
	R7off	1,390	0,748					
Pair 8	**R8on**	2,894	0,593	0,080	11,048	0,000		
	R8off	0,955	0,652					
Pair 9	**R9on**	2,975	0,537	0,117	9,616	0,000		
	R9off	1,035	0,843					
Pair 10	**R10on**	2,916	0,508	0,074	11,289	0,000		
	R10off	0,697	0,711					

The Shapiro-Wilk Test shows that the assumption of normality has not been violated (p>0.05) except for Run 7 (p<0.05) In this regard, we conduct a Wilcoxon Test for Run 7 and a dependent T-Test for the rest of the runs. Results for the dependent T-Test determine that there is a significant difference in the means of fixations per second between OnRun and OffRun periods for run 3 onwards (t>0, p<0.05). The Wilcoxon Test shows that for Run 7 there is also a statistically significant difference (Z=-3.848, p<0.05).

The direction of our t-values for the dependent T-Test (t>0), together with the direction of the Z-value of the Wilcoxon Test (Z<0), indicate that there is a higher number of fixations in OnRun periods than in OffRun periods. Like for condition *shape*, this result is in line with the necessity to focus more on different items when participants are looking for one specific item than in the resting state that comes once they have successfully found the target item.

As in condition *shape*, we have found statistically significant differences in the number of fixations per second from Run 3 onwards. This fact indicates that the task induces a level of cognition high enough to be detected through the analysis of fixations. In addition, it suggests that the level of cognition induced is higher than for condition *colour*.

In order to understand whether there are statistically significant differences between the mean number of fixations per second in our three conditions, *colour*, *shape* and *colour&shape*, we can use a repeated measures ANOVA test. To be able to conduct this test, our data has to fulfill the assumption of sphericity, tested through the Mauchly's Test of Sphericity and whose results can be seen in Table 5.

Table 5. Results to test for sphericity for *fixations* in the data for our three conditions – *colour*, *shape* and *colour&shape*-.

Measure: Fixations

Mauchly's Test of Sphericity

Within-Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Condition	0,912	2,314	2	0,315	0,919	0,985	0,500

Observing the results presented, Mauchly's Test of Sphericity indicates that the assumption of sphericity has not been violated, (χ^2 (2) = 2.314, p = 0.315) and therefore we can test our data for differences between conditions through a repeated measures ANOVA test. Table 6 presents descriptive statistics about the total number of fixations per second (OnRun period + OffRun period) for our three conditions that are important to analyze our data. Additionally, Table 7 presents the results of a repeated measures ANOVA test.

Table 6. Descriptive statistics of the total number of *fixations* per second for our three conditions (*colour, shape* and *colour&shape).*

Descriptive Statistics

	Mean (NºFix/s)	Std. Deviation	N
1 - Colour	2,0309	0,2005	27
2 - Shape	2,2274	0,5533	27
3 - C&S	2,6674	0,4117	27

Table 7. Results of a repeated measures ANOVA test for our three conditions – *colour*, *shape* and *colour&shape*- and measurement *fixations.*

Measure: Fixations

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Condition	Sphericity Assumed	5,736	2	2,868	20,948	0,000	0,446
Error(condition)	Sphericity Assumed	7,119	52	0,137			

A repeated measures ANOVA determined that the number of fixations per second differed statistically significantly between our three conditions (F (2, 7.119) = 20.948, $p < 0.05$). However, this test informs us that we have an overall significant difference in means, but we do not know where those differences occurred.

The following table, Table 8, presents the results of our pairwise comparison with a Bonferroni correction, which allows us to specifically discover which conditions differ. This test compares the differences in the mean number of fixations per second between conditions at a 95% confidence interval divided by the number of pairwise comparisons that, in our case, is 3.

For our three conditions, firstly it tests one condition, says *colour*, against each of the other two, *shape* and *colour&shape*. Afterwards, it tests the second condition, says *shape*, against each of the other two, *shape* and *colour&shape.* Finally, it tests the last condition, *colour&shape* against the former two, *colour* and *shape*.

Table 8. Results from the ANOVA test using a Bonferroni correction for the measurement *fixations*.

Measure: Fixations

Pairwise Comparisons

(I) condition		Mean Difference (I-J) (NºFix/s)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
Colour	Shape	-0,196	0,109	0,251	-0,476	0,083
	C&S	-,636	0,085	0,000	-0,853	-0,420
Shape	Colour	0,196	0,109	0,251	-0,083	0,476
	C&S	-,440	0,107	0,001	-0,713	-0,167
C&S	Colour	,636	0,085	0,000	0,420	0,853
	Shape	,440	0,107	0,001	0,167	0,713

Our repeated measures ANOVA test using the Bonferroni correction has revealed that there is a slight increment in the number of fixations per second from condition *colour* to condition *shape* (2.03 ±0.20 fix/s vs 2.23 ± 0.55 fix/s), which was not statistically significant (p=0.251). However, the number of fixations per second measured for condition *colour&shape* indicates a higher increment (2.67 ± 0.41 fix/s), which differs statistically from conditions *colour* (p = 0.000) and *shape* (p = 0.001).

In this regard, we can conclude that a higher cognitive load can be represented by a significant increment in the number of fixations per second, being condition *colour&shape* the most cognitively demanding.

8.2 Saccades

Saccades refer to the period of time where the eye shifts from a focused point (fixation) to another. Based on Chen et al. (2011), García Barrios et al. (2004) and Rudmann et al. (2003)'s conclusions described in Chapter 3.¸ one can expect a direct relation between saccades' count and cognitive load, i.e. an increment in the number of saccades for a higher level of cognition. But, does saccade count increase as the difficulty of the task increases? Is there a difference in saccade count during periods of higher level of cognition in comparison to more relaxed states? Can saccades be used as an indicator to measure cognitive load?

Through the results presented in this section, we try to find out the answer to those questions by studying if and how the number of saccades per second changes between OnRun and OffRun periods and if an active visual search task induces more or less cognitive load depending on the condition that describes the task – *colour*, *shape*, or *colour&shape.*

Colour

> We use a dependent T-Test to find out if there are differences in the mean values of saccades per second when participants visually search for the element target in comparison to the following period, once the target has been found. Additionally, normality has been tested using a Shapiro-Wilk test. For periods of OnRun-Off Run, for which the data does not follow a normal distribution, a Wilcoxon Test has been utilized. The results of the indicated tests, averaged by participants for each of the runs, are presented in Table 9.

Table 9. Results of the Shapiro-Wilk test, dependent T-Test and Wilcoxon test for the gaze event *saccades* for condition *colour* averaged by participant, for each of OnRun/OffRun periods.

NºSacc/s Colour		**Descriptive Statistics**		**Shapiro-Wilk Test**	**Dependent T-Test**		**Wilcoxon Test**	
		Mean Sacc/s	Std. Deviation	p-value	t-value	p-value	z-value	p-value
Pair 1	**R1on**	1,764	0,382	0,464	-0,312	0,758		
	R1off	1,808	0,655					
Pair 2	**R2on**	1,614	0,351	0,433	0,651	0,521		
	R2off	1,551	0,571					
Pair 3	**R3on**	1,678	0,289	0,613	1,105	0,279		
	R3off	1,582	0,513					
Pair 4	**R4on**	1,776	0,453	0,693	1,110	0,277		
	R4off	1,622	0,652					
Pair 5	**R5on**	1,807	0,402	0,544	0,934	0,359		
	R5off	1,705	0,528					
Pair 6	**R6on**	1,818	0,375	0,569	2,026	0,053		
	R6off	1,611	0,468					
Pair 7	**R7on**	1,929	0,424	0,000			-1,802	0,072
	R7off	1,819	1,168					
Pair 8	**R8on**	1,922	0,386	0,491	2,442	0,022		
	R8off	1,617	0,529					
Pair 9	**R9on**	1,938	0,528	0,449	3,139	0,004		
	R9off	1,510	0,575					
Pair 10	**R10on**	2,032	0,543	0,013			-2,691	0,007
	R10off	1,563	0,603					

Observing Table 9, a Shapiro-Wilk Test carried out in our data exposes that, except for Run 7 and Run 10 ($p < 0.05$), data for the rest of the pairs OnRun/OffRun follows a normal distribution ($p > 0.05$).

For data that is normally distributed (green highlighting), we have tested for significant differences in the mean values of saccades per second through a dependent T-test. Results presented in the table above reveal that there are no significant differences (red highlighting) between the number of saccades per second for OnRun and OffRun pairs below Run 6 ($p > 0.05$). Additionally and only for Run 1, the t-value indicates that the number of saccades is greater in the OffRun period than in the OnRun one (t-value < 0), what is an indicator that the level of cognition is not sufficiently high.

Furthermore, results obtained from the Wilcoxon Test for Run 7 demonstrate that OnRun and OffRun periods present no significant differences in the number of saccades per second, although the number of saccades is greater while participants search for the element (OnRun) (z-value=-1.802; $p > 0.05$).

However, unlike fixations, one can already observe a statistically significant difference in the mean number of saccades per second in pairs from Run 8 on, for both the dependent T-Test and Wilcoxon Test ($p < 0.05$). Moreover, the results of the t- and z-values suggest that the number of saccades per second is greater in OnRun periods than in OffRun periods.

Shape

For condition *shape* we have followed the same procedure as for condition *colour*. Results obtained for Shapiro-Wilk test, dependent T-Test and Wilcoxon Test can be seen in Table 10.

Table 10. Results of the Shapiro-Wilk test, dependent T-Test and Wilcoxon test, for the gaze event *saccades* for condition *shape* averaged by participant, for each of OnRun/OffRun periods.

N°Sacc/s - Shape		**Descriptive Statistics**		**Shapiro -Wilk Test**	**Dependent T-Test**		**Wilcoxon Test**	
		Mean Sacc/s	Std. Deviation	p-value	t-value	p-value	Z-value	p-value
Pair 1	**R1on**	2,247	0,791	0,206	2,858	0,008		
	R1off	1,834	0,668					
Pair 2	**R2on**	1,929	0,751	0,824	2,965	0,006		
	R2off	1,584	0,598					
Pair 3	**R3on**	1,827	0,655	0,804	2,197	0,037		
	R3off	1,554	0,699					
Pair 4	**R4on**	2,127	0,764	0,024			-3,700	0,000
	R4off	1,442	0,736					
Pair 5	**R5on**	2,230	0,693	0,214	4,525	0,000		
	R5off	1,454	0,877					
Pair 6	**R6on**	2,109	0,762	0,693	3,579	0,001		
	R6off	1,482	0,783					
Pair 7	**R7on**	2,050	0,803	0,028			-2,787	0,005
	R7off	1,435	0,689					
Pair 8	**R8on**	2,114	0,714	0,630	4,972	0,000		
	R8off	1,344	0,730					
Pair 9	**R9on**	2,196	0,653	0,025			-4,349	0,000
	R9off	1,362	0,586					
Pair 10	**R10on**	2,025	0,601	0,598	4,317	0,000		
	R10off	1,459	0,582					

The Shapiro-Wilk test shows that for pairs 1 to 3, 5, 6, 8 and 10 there is no significant difference between the mean saccades per second in OnRun and OffRun periods, indicating that the data is normally distributed (p > 0.05, highlighted in green). Fulfilling the assumption of normality, we check for significant differences in the data through a dependent T-Test. For data that is not normally distributed (pairs 4, 7 and 9; $p < 0.05$) a Wilcoxon signed-rank test is applied.

Both dependent T-Test and Wilcoxon test show statistically significant differences in the mean number of saccades per second comparing OnRun and OffRun periods ($p < 0.05$). Moreover, one can appreciate an increment in t- and z-values in comparison to condition *colour.* The direction of t-values for the dependent T-Test, together with the z-values for the Wilcoxon signed-rank test ($t > 0$; $z < 0$) indicate that the number of saccades per second is larger in OnRun periods than in OffRun ones, following participants' necessity to shift their gaze a higher number of times when searching for a target than once they have found the target item.

Furthermore, there is a slight increment in the t-values from run 1 to run 10, which represents a greater difference in the mean values between OnRun and OffRun periods. This increment is in line with the increment in the number of distractors.

Colour&Shape

As well as in conditions *colour* and *shape,* data for condition *colour&shape* has been tested for significant differences between OnRun and OffRun periods. In order to decide whether to use a dependent T-Test or its non-parametric version, a Wilcoxon signed-rank test, data has to be normally distributed. Conducting a Shapiro-Wilk test gives us that information about the distribution of the data. In the table below, Table 11, results for all three tests are presented.

Table 11. Results of the Shapiro-Wilk test, dependent T-Test and Wilcoxon test for the gaze event *saccades* for condition *colour&shape* averaged by participant, for each of OnRun/OffRun periods.

NºSacc/s - Colour&Shape		Descriptive Statistics		Shapiro -Wilk Test	Dependent T-Test		Wilcoxon Test	
		Mean Sacc/s	Std. Deviation	p-value	t-value	p-value	Z-value	p-value
Pair 1	**R1on**	1,978	0,678	0,959	2,900	0,008		
	R1off	1,638	0,638					
Pair 2	**R2on**	2,046	0,490	0,031			-2,502	0,012
	R2off	1,752	0,574					
Pair 3	**R3on**	2,459	0,721	0,000			-3,391	0,001
	R3off	1,756	1,537					
Pair 4	**R4on**	2,389	0,572	0,223	6,083	0,000		
	R4off	1,502	0,544					
Pair 5	**R5on**	2,563	0,657	0,593	7,969	0,000		
	R5off	1,062	0,707					
Pair 6	**R6on**	2,766	0,699	0,823	10,166	0,000		
	R6off	1,023	0,542					
Pair 7	**R7on**	2,676	0,625	0,413	8,628	0,000		
	R7off	1,128	0,662					
Pair 8	**R8on**	2,779	0,640	0,173	11,959	0,000		
	R8off	0,816	0,574					
Pair 9	**R9on**	2,845	0,622	0,086	12,017	0,000		
	R9off	0,786	0,583					
Pair 10	**R10on**	2,801	0,581	0,187	11,705	0,000		
	R10off	0,625	0,641					

The Shapiro-Wilk Test announces that the assumption of normality has been violated for runs 2 and 3, ($p < 0.05$), while data for the rest of the pairs is normally distributed ($p > 0.05$, green highlighting). In this regard, we cannot conduct a dependent T-Test for Run 2 and 3, since the assumption of normality is not fulfilled. For those pairs, we conduct a Wilcoxon signed-rank Test and we carry out a dependent T-Test for the rest of the runs.

Results for both tests determine that there are significant differences in the means of saccades per second between OnRun and OffRun periods for all ten runs ($p < 0.05$).

Furthermore, the t-values of the dependent T-Test ($t > 0$) and the z-values of the Wilcoxon test ($z < 0$), indicate that there is a significant difference in the mean values for OnRun and OffRun periods, being the number of saccades per second higher in OnRun periods.

It is interesting to point out the increment of t- and z- values from low values in earlier runs to higher values in later runs. This finding is in consonance with the increment in the number of distractors per run, as participants need to shift their gaze more often when there are more distractors present, in order to cover all of them and discard them if they are not the target item.

Comparing these values with the t- and z-values from conditions *colour* and *shape,* one can appreciate a great increment in those values, which is in line with the increment in level of cognition imposed by a condition where participants, in order to find a target item, have not only to focus on one property but in two, colour and shape.

In order to answer the question about if saccades can be used as an indicator of cognitive load, we have to study whether there are significant differences between the mean number of saccades per second for our three conditions, *colour*, *shape* and *colour&shape*. In this regard, we can conduct a repeated measures ANOVA test if our data achieves the assumption of sphericity, which can be tested through Mauchly's Test of Sphericity. Results for this test can be seen in Table 12.

Table 12. Results to test for sphericity in the data for *saccades* and for our three conditions – *colour*, *shape* and *colour&shape*-.

Measure: Saccades

Mauchly's Test of Sphericity

Within-Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Condition	0,912	2,313	2	0,315	0,919	0,985	0,500

Results from Maulchy's Test of Sphericity shown in the table above indicate that the assumption of sphericity is not violated (χ^2 (2) = 2.313, p = 0.315) and consequently, our data can be tested using a repeated measures ANOVA test.

Additionally, we calculate the mean values in the number of saccades per second and per condition, what implicates that the number of saccades per second represents the sum of both OnRun and OffRun periods. These results are presented in Table 13. The results of the ANOVA test can be seen in Table 14.

Table 13. Descriptive statistics of the total number of *saccades* per second for our three conditions (*colour, shape* and *colour&shape).*

Descriptive Statistics

	Mean (NºSac/s)	Std. Deviation	N
1 - Colour	1,8279	0,2644	27
2 - Shape	2,0855	0,6234	27
3 - C&S	2,5303	0,4846	27

Table 14. Results of a repeated measures ANOVA test for *saccades* for our three conditions – *colour*, *shape* and *colour&shape*–.

Measure: Nº Sac/s **Tests of Within-Subjects Effects**

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Condition	Sphericity Assumed	6,820	2	3,410	19,363	0,000	0,427
Error (condition)	Sphericity Assumed	9,157	52	0,176			

Results obtained from the repeated measures ANOVA test determine a statistically significant difference between the number of saccades per second among the conditions (F (2, 6.820) = 19.363; $p < 0.05$).

In order to locate the differences in the mean number of saccades per second between our three conditions, we conduct a pairwise comparison with a Bonferroni correction. The results of this test can be seen in Table 15.

Table 15. Results from the ANOVA test using a Bonferroni correction for the measurement *saccades*.

Measure: Saccades **Pairwise Comparisons**

(I) condition		Mean Difference (I-J) (NºSac/s)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
Colour	Shape	-0,258	0,124	0,142	-0,574	0,059
	C&S	-,702	0,096	0,000	-0,948	-0,457
Shape	Colour	0,258	0,124	0,142	-0,059	0,574
	C&S	-,445	0,121	0,003	-0,754	-0,135
C&S	Colour	,702	0,096	0,000	0,457	0,948
	Shape	,445	0,121	0,003	0,135	0,754

One can observe that there is a slight increment in the number of saccades per second between condition *colour* and condition *shape* (from 1.83 ± 0.26 sac/s to 2.09 ± 0.62 sac/s) although this difference is not statistically significant (p = 0.142). Notwithstanding, we can appreciate a greater increment in the number of saccades per second for condition *colour&shape* (2.53 ± 0.48 sac/s) that differ statistically significantly from conditions *colour* (p = 0.000) and condition *shape* (p = 0.003). Considering these results, one can conclude that a higher level of cognition induces an increment in the number of saccades per second.

Through the analysis of saccades, we can conclude that they are a valid indicator to measure cognitive load and they can detect the cognitive load in earlier states than fixations, as shown in the condition *colour*.

8.3 Blinks

Some researchers relate eye blinks to user's states of attention and cognitive load. Chen et al. (2011) and García Barrios et al. (2004) observed that there is an indirect relation between blink rate and the level of cognition, being the higher the load, the lower the blink rate. Results from our pre-study were not significant enough to draw conclusions due to tasks' shortness, which implied not having enough blinks to analyze them properly in relation to cognitive load. Increasing the task duration, we have been able to collect a sufficient number of blinks allowing us to ponder questions such as: can the analysis of blinks help to get a deeper understanding about cognitive load? Do blinks decrease as the level of cognition imposed by the difficulty of the task increases? Is there a difference in the blink rate for periods OnRun and OffRun?

In order to answer these questions, we study the blinking rate for our three conditions that describe the task – *colour*, *shape*, or *colour&shape*.

Colour

In order to study the distribution of our data, we make use of a Shapiro-Wilk test. If our data is normally distributed, we use a dependent T-Test to look for differences in blink rate between OnRun and OffRun periods. If our data violates the assumption of normality necessary for a T-Test, we test for significant differences through a Wilcoxon signed-rank test. The results of these tests averaged by participants and runs can be seen in Table 16.

Table 16. Results of the Shapiro-Wilk test, dependent T-Test and Wilcoxon test, and statistical differences in the gaze event *blinks* for condition *colour* averaged by participant, for each of OnRun/OffRun periods.

N°Blink/s Colour		**Descriptive Statistics**		**Shapiro-Wilk Test**	**Dependent T-Test**		**Wilcoxon Test**	
		Mean (blink/s)	Std. Deviation	p-value	t-value	p-value	z-value	p-value
Pair 1	**R1on**	1,764	0,382	0,000			-,649	0,517
	R1off	1,808	0,655					
Pair 2	**R2on**	1,614	0,351	0,000			-,288	0,773
	R2off	1,551	0,571					
Pair 3	**R3on**	1,678	0,289	0,030			-,889	0,374
	R3off	1,582	0,513					
Pair 4	**R4on**	1,776	0,453	0,027			-1,273	0,203
	R4off	1,622	0,652					
Pair 5	**R5on**	1,807	0,402	0,018			-,985	0,325
	R5off	1,705	0,528					
Pair 6	**R6on**	1,818	0,375	0,003			-2,042	0,041
	R6off	1,611	0,468					
Pair 7	**R7on**	1,929	0,424	0,005			-1,802	0,072
	R7off	1,819	1,168					
Pair 8	**R8on**	1,922	0,386	0,152	2,442	0,022		
	R8off	1,617	0,529					
Pair 9	**R9on**	1,938	0,528	0,002			-2,859	0,004
	R9off	1,510	0,575					
Pair 10	**R10on**	2,032	0,543	0,000			-2,691	0,007
	R10off	1,563	0,603					

Results from the Shapiro-Wilk Test reveal that the data does not follow a normal distribution ($p < 0.05$) except for Pair 8, which is normally distributed ($p = 0.152$).

For data that violates the assumption of normality, we have conducted a non-parametric Wilcoxon test and its results show no significant difference in the blink rate for periods OnRun and OffRun for runs 1 to 5 and for Run 7 ($p > 0.05$). Additionally, Wilcoxon test and dependent T-Test indicate that from Run 8 on the number of blinks per second differs significantly from OnRun/OffRun periods ($p < 0.05$).

Observing the mean values for each run, one can appreciate a slight increment in the number of blinks per second for OnRun periods in which participants search for the target item in comparison to OffRun periods. These results for condition *colour* contradict the findings done by Chen et al. (2011) and García Barrios et al. (2004), who claimed that there is an indirect relation between blink rate and the level of cognition. It is interesting to study if our results apply to all three conditions or if, on the contrary, they are isolated to condition *colour,* as this condition offers the fewest level of cognition, which could not be identifiable relying only on the blink rate.

Shape

The blink rate for condition *shape* has been analyzed following the same procedure as for condition *colour*. Firstly, we have tested the data for normality through a Shapiro-Wilk test. In order to find if there are significant differences between OnRun and OffRun periods, we test our pairs through a Wilcoxon signed-rank test, indicated for data that is not normally distributed. The results of these tests are presented in Table 17.

Table 17. Results of the Shapiro-Wilk test and Wilcoxon test for the gaze event *blinks* in condition *shape* averaged by participant, for each of OnRun/OffRun periods.

NºBlink/s Shape		**Descriptive Statistics**		**Shapiro -Wilk Test**	**Wilcoxon Test**	
		Mean (blinks/s)	Std. Deviation	p-value	Z-value	p-value
Pair 1	**R1on**	0,074	0,086	0,000	-4,432	0,000
	R1off	0,367	0,300			
Pair 2	**R2on**	0,094	0,110	0,019	-4,305	0,000
	R2off	0,378	0,330			
Pair 3	**R3on**	0,102	0,117	0,001	-4,445	0,000
	R3off	0,329	0,251			
Pair 4	**R4on**	0,090	0,098	0,004	-4,012	0,000
	R4off	0,328	0,298			
Pair 5	**R5on**	0,092	0,093	0,000	-4,349	0,000
	R5off	0,330	0,296			
Pair 6	**R6on**	0,114	0,111	0,000	-4,276	0,000
	R6off	0,463	0,509			
Pair 7	**R7on**	0,108	0,116	0,002	-4,397	0,000
	R7off	0,369	0,288			
Pair 8	**R8on**	0,106	0,124	0,001	-4,280	0,000
	R8off	0,346	0,295			
Pair 9	**R9on**	0,102	0,106	0,002	-4,036	0,000
	R9off	0,344	0,301			
Pair 10	**R10on**	0,107	0,116	0,000	-4,517	0,000
	R10off	0,462	0,389			

We can observe from the results obtained from the Shapiro-Wilk Test, that the data does not follow a normal distribution ($p < 0.05$). In this regard, we utilize a Wilcoxon test, which does not require the data to be normally distributed. The results of the test for all runs show significant differences in the number of blinks per second between OnRun and OffRun periods ($z < 0$, $p < 0.05$).

Unlike condition *colour*, for condition *shape* we can appreciate a decrement in the number of blinks per second for OnRun periods, attending to the necessity of the participant to focus more on discovering the target item and trying to miss the less possible by blinking.

These results for condition *shape* are in line with the findings made by Chen et al. (2011) and García Barrios et al. (2004) and in line with an increment in the level of cognition imposed by an increment in the difficulty of the task.

Colour&Shape

Results for condition *colour&shape* are presented in Table 18. In the same way as for conditions *colour* and *shape,* we have tested the distribution of our data through a Shapiro-Wilk test, and differences in the mean number of blinks per second for OnRun/OffRun periods are detected through a dependent T-Test (in data that is normally distributed) and through a Wilcoxon signed-rank test (for data that does not follow a normal distribution).

Table 18. Results of the Shapiro-Wilk test, dependent T-Test and Wilcoxon test for the gaze event *blinks* for condition *colour&shape* averaged by participant, for each of OnRun/OffRun periods.

NºBlink/s Colour&Shape		**Descriptive Statistics**		**Shapiro-Wilk Test**	**Dependent T-Test**		**Wilcoxon Test**	
		Mean blink/s	Std. Deviation	p-value	t-value	p-value	Z-value	p-value
Pair 1	**R1on**	0,160	0,158	0,000			-3,943	0,000
	R1off	0,397	0,367					
Pair 2	**R2on**	0,157	0,127	0,004			-4,023	0,000
	R2off	0,379	0,285					
Pair 3	**R3on**	0,107	0,106	0,000			-4,457	0,000
	R3off	0,388	0,258					
Pair 4	**R4on**	0,143	0,121	0,000			-4,381	0,000
	R4off	0,458	0,541					
Pair 5	**R5on**	0,119	0,134	0,198	-4,069	0,000		
	R5off	0,295	0,244					
Pair 6	**R6on**	0,111	0,118	0,000			-3,700	0,000
	R6off	0,371	0,500					
Pair 7	**R7on**	0,117	0,114	0,005			-3,404	0,001
	R7off	0,346	0,362					
Pair 8	**R8on**	0,098	0,103	0,005			-2,005	0,045
	R8off	0,205	0,245					
Pair 9	**R9on**	0,116	0,108	0,003			-2,400	0,016
	R9off	0,236	0,275					
Pair 10	**R10on**	0,100	0,094	0,001			-1,857	0,007
	R10off	0,166	0,228					

The Shapiro-Wilk test indicates that the data, but for Run 5 ($p > 0.05$), is not normally distributed ($p < 0.05$). Data that does follow a normal distribution (Pair 5) has been tested for statistical differences in the mean number of blinks per second through a dependent T-Test. A Wilcoxon signed-rank test is applied to data that is not normally distributed.

Observing the table above, we can detect, for both tests, the mean differences in number of blinks per second for OnRun/OffRun periods differ significantly ($p < 0.05$; $t < 0$; $z < 0$;).

As for condition *shape*, for condition *colour&shape* there is also a decrement in the number of blinks per second in OnRun periods, in relation to an increment in the level of cognition.

Attending to the differences found in the blink rate in OnRun and OffRun periods, it is interesting to investigate how these differences affect our three conditions. Is there a statistically significant difference in the blink rate between condition *colour* and condition *shape* or *colour&shape*? Or is there a difference between conditions *shape* and *colour&shape*?

In this regard, we conduct a repeated measures ANOVA test for data that achieves sphericity, which is tested through the Mauchly's Test of Sphericity. In Table 19 the results of this test are presented.

Table 19. Results to test for sphericity in the data for our three conditions – *colour*, *shape* and *colour&shape*- for the measurement *blinks*.

Measure: Blinks

Mauchly's Test of Sphericity

Within-Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Condition	0,817	5,052	2	0,080	0,845	0,898	0,500

Results from Maulchy's Test of Sphericity show that the assumption of sphericity is not violated (χ^2 (2) = 5.052, p = 0.080) and consequently, our data can be tested using a repeated measures ANOVA test.

Furthermore, Table 20 presents descriptive statistics about the total number of blinks per second per task (OnRun blink/s + OffRun blink/s).

Table 20. Descriptive statistics about the total number of *blinks* per second for our three conditions (*colour, shape* and *colour&shape).*

Descriptive Statistics

	Mean (NºBlink/s)	Std. Deviation	N
1 - Colour	0,1393	0,1204	27
2 - Shape	0,0988	0,0916	27
3 - C&S	0,1227	0,0962	27

In order to find out if there are significant differences in the mean number of blinks per second between our three conditions, we utilize a repeated measures ANOVA test, whose results are summarized below in Table 21.

Table 21. Results of a repeated measures ANOVA test for *blinks* for our three conditions – *colour*, *shape* and *colour&shape-*.

Measure: Blinks

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Condition	Sphericity Assumed	0,022	2	0,011	6,311	0,004	0,195
Error (condition)	Sphericity Assumed	0,092	52	0,002			

The repeated measures ANOVA test's results confirm that there is a statistically significant difference in the blink rate between our three conditions (F (2, 0.092) = 6.311; p = 0.004). We conduct a pairwise comparison with a Bonferroni correction to get a deeper understanding about between which specific conditions the significant differences arise. Table 22 presents these results.

Table 22. Results from the ANOVA test using a Bonferroni correction for the measurement *blinks* in order to find out where the differences between our conditions occur.

Measure: Blinks

Pairwise Comparisons

(I) condition		Mean Difference (I-J) (NºBlink/s)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
Colour	Shape	0,041	0,013	0,017	0,006	0,075
	C&S	0,017	0,011	0,465	-0,012	0,046
Shape	Colour	-0,041	0,013	0,017	-0,075	-0,006
	C&S	-0,024	0,009	0,045	-0,047	0,000
C&S	Colour	-0,017	0,011	0,465	-0,046	0,012
	Shape	0,024	0,009	0,045	0,000	0,047

From the information presented above, one can observe that there is a slight decrement in the number of blinks per second between condition *colour* and condition *shape* (from 0.139 ± 0.12 blink/s to 0.099 ± 0.09 blink/s) that is statistically significant (p = 0.017).

Notwithstanding, we can appreciate an increment in the blink rate for condition *colour&shape* (0.123 ± 0.096 blinks/s) respect to condition *shape* that differs statistically significantly (p = 0.045) and a decrement in the blink rate respect to condition *colour* that is, however, not statistically significant (p = 0.465). We expected *colour&shape* to be statistically different to condition *colour*, in line with the measurements *fixations* and *saccades*. However, the results contradict our expectations. A possible explanation will be discussed in the next chapter.

8.4 Pupil dilation

The pupillary response is another widely studied eye event. Pupil size varies, in principle, to adapt the eye to changes in the luminance, contracting when the environment becomes brighter and dilating when the environment becomes darker, in order to acquire more light.

Nonetheless, several studies relate pupil dilation to levels of cognition and emotional stimuli (Porta et al. (2012)), such as tiredness. Experiments realized by Chen et al. (2011), Rafiqi et al. (2015) and Rudmann et al. (2003) show a direct relation between the pupil size and the level of cognition. They observed an increment in pupil diameter as a response to a high cognitive state.

In this regard, it is interesting to study how pupil dilation indicates which condition produces a higher cognitive load in our visual search tasks. Beyond study changes in the pupil diameter at a task- or even run-level, it is interesting to determinate changes in the pupil size at a lap-level. By doing so, one can get an accurate idea about how pupil fluctuation evolves across time. It is relevant to consider whether the number of distractors influences the level of cognitive load. Does the cognitive load increase with increasing number of distractors? How does pupil diameter evolve in OnRun periods compared to OffRun periods? Are there differences in the evolution of pupil size across conditions?

The outcome of our experiment is presented by condition, *colour, shape* and *colour&shape*.

Colour

Figure 23 represents the fluctuation in pupil size averaged by participants for each 30 laps. Each division represents one lap of pupil dilation for OnRun and OffRun periods.

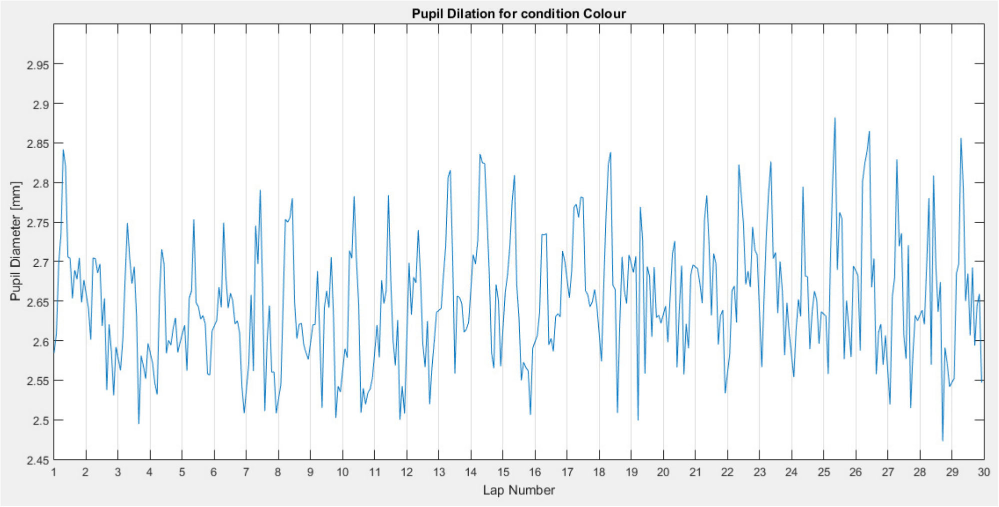


Figure 23. Average pupil dilatation fluctuation for condition *colour*.

At the beginning of the task, pupil dilatation presents a peak of increment as the participant is in an "alert state" waiting for the beginning of the task and trying to perform it as well as possible. This fact is translated into a higher cognitive state. From the second lap on, the pupil diameter decreases what is an indicator that the cognitive load decreases too once the participant gets to know the dynamic of the task.

Figure 24 presents a detailed pupil fluctuation. For each of the laps, one can clearly observe how at the beginning of the lap (A point) the pupil size starts increasing as the participant visually searches for the object (OnRun). It reaches its maximum dilation peak when the participant finds the object (B point). From that moment on (OffRun), the participant does not need to do anything but to wait for the next round to start (C point). His relaxation is reflected in the decrement of his level of cognition, in consonance with a decrement in his pupil diameter. This sequence is repeated for each of the laps.

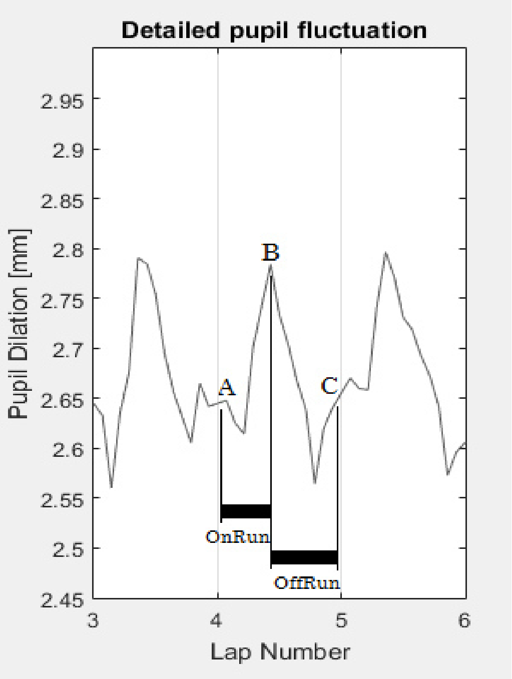


Figure 24. Detailed pupil size fluctuation for OnRun and OffRun periods.

However, in Figure 23 one cannot appreciate a general increment in the level of cognition in relation to an increment in the number of distractors towards the end of the task.

Shape

Figure 25 represents the pupil fluctuation for condition *shape*. Likewise, condition *colour*, one can appreciate the increment in pupil dilation for OnRun periods and the pupil dilation decrement for OffRun periods, in consonance with the fluctuation in the participant's level of cognition.

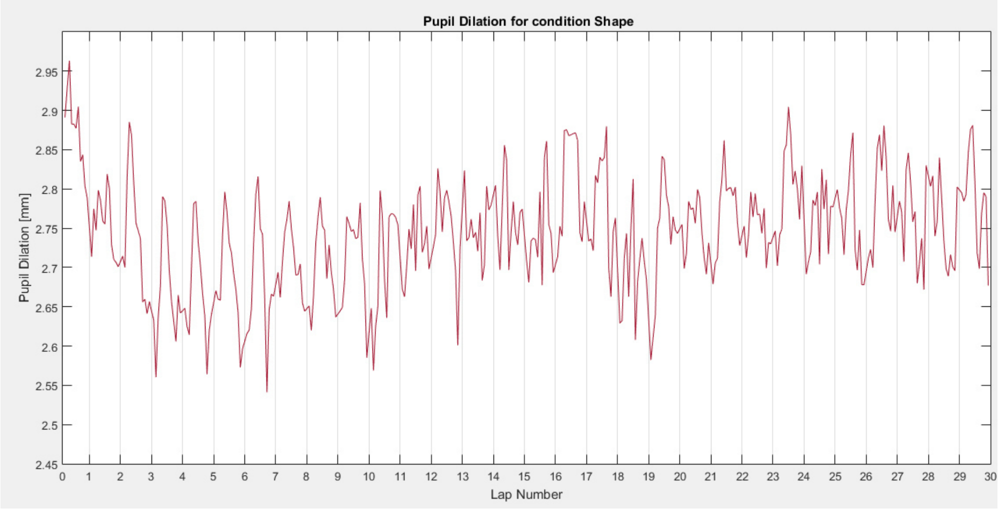


Figure 25. Average pupil dilatation variation for the condition *shape*.

Furthermore, towards the end of the task, variations between the minimum and maximum pupil diameter within a lap are significantly shorter than at the beginning of the task. This phenomenon could be related to the overall level of cognition that is higher at the end of the task, imposed by an increment in the difficulty as participants need to find a target item among a greater number of distractors.

Colour&Shape

The fluctuation in pupil size for condition *colour&shape* is presented in Figure 26. As for condition *colour* and condition *shape,* the fluctuation in pupil size from OnRun periods to OffRun periods is clearly identifiable.

Furthermore, the absolute dilation per lap (maximum peak – minimum peak) is significantly smaller for condition *colour&shape* in comparison to conditions *colour* and *shape.*

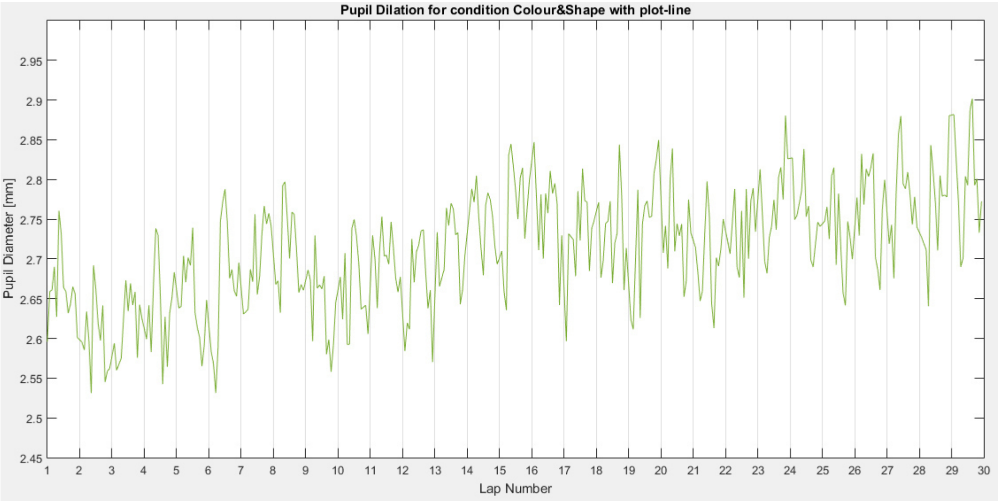


Figure 26. Average pupil dilatation variation for the condition *colour&shape.*

One of the reasons to study how the pupil diameter evolves in a lap level was to find out if there exist a relation between cognitive load and difficulty of the task, imposed by the increasing number of distractors per run. According to these results, one cannot observe a significant increment in the pupil diameter when the number of distractors increases for conditions *colour* and *shape* (Figure 23 and 24 respectively).

However, as shown in Figure 26, one can observe a significant increment in the pupil dilation towards the end of the task, revealing an increment in the level of cognition. This increment is in line with the augmentation in the number of distractors. Figure 27 presents the same pupil dilation as Figure 26 but it overlays a plot-line to indicate the increment in pupil size. These results are in line with a higher level of cognition imposed by the increment in difficulty for condition *colour&shape*.

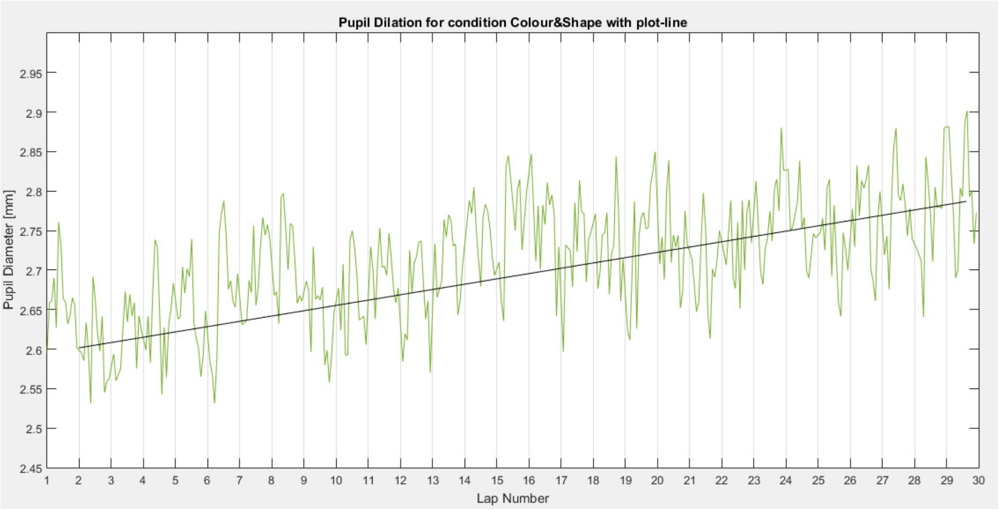


Figure 27. Average pupil dilatation increment indicated by a plot-line, in relation to an increment in the number of distractors for the condition *colour&shape*.

Additionally, we ponder the question about if there are differences in the evolution of pupil size across conditions and in an affirmative case, which are those differences.

In this regard, to study how the level of cognition is affected by each of the conditions, *colour, shape* and *colour&shape*, we present a combination of the pupil diameter fluctuation for the three conditions. Figure 28 presents a plot for each of the conditions differentiated its colour and for each of the thirty laps. A bigger chart of Figure 28 can be found in Appendix B.

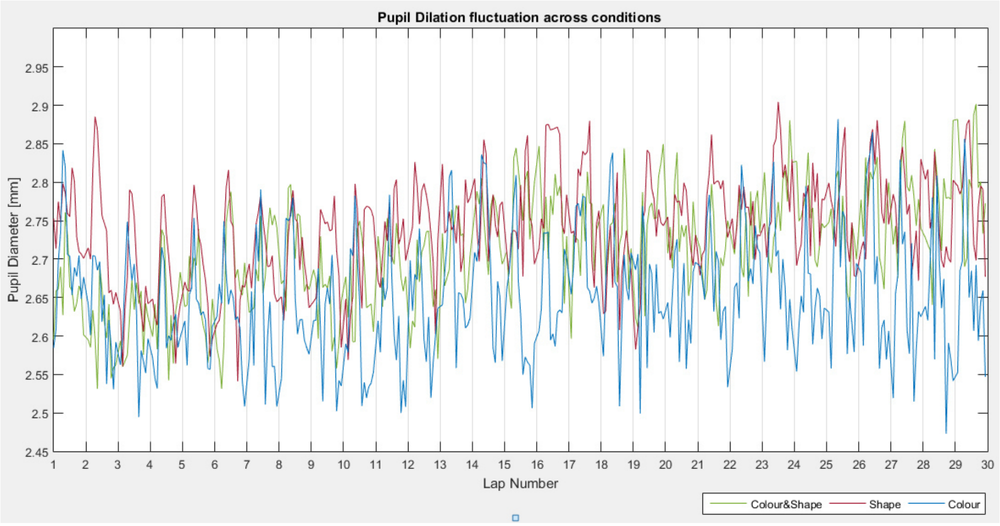


Figure 28. Average pupil dilatation for the conditions *colour, shape* and *colour&shape.*

From the results presented on the graph, one can observe that there are differences in pupil dilation across conditions. Condition *colour* presents the lowest minimum pupil diameter on average for all laps. This is an indicator that condition *colour* is the one that imposes the fewest cognitive load. For conditions *shape* and *colour&shape,* the difference is not that pronounced, although condition *colour&shape* presents a lower pupil diameter average at the beginning of the task that increases towards the end of the task more pronouncedly than condition *shape,* in line with the level of cognition, which is the highest for condition *colour&shape.*

Due to the high dependence between the pupil dilation and the participant, together with the continuous nature in the representation of the data, we have not performed any statistical analysis.

In conclusion, pupil dilation is a good indicator to detect changes in the level of cognition at a lap level, although averaging the pupil dilation over the entire task does not make sense, since periods of higher level of cognition negates periods of lower cognitive load and therefore may not show significant increment. Further conclusions will be offered in the discussion chapter, Chapter 9.

8.5 NASA TLX

This section presents the NASA TLX results obtained in a more indirect and subjective manner from the questionnaires that participants filled right after the completion of each task. We use these findings in order to complement the eye tracking results obtained and discussed in previous sections of this chapter.

The NASA TLX questionnaire provides an overall workload score based on ratings on six subscales: *mental demand, physical demand, temporal demand, performance, effort* and *frustration.*

These subscales are rated on a scale from 0 to 100-points in increments of 5. The overall workload score is achieved as the average workload score for all six subscales. Figure 29 presents a chart with the mean score per subscale for all three conditions.

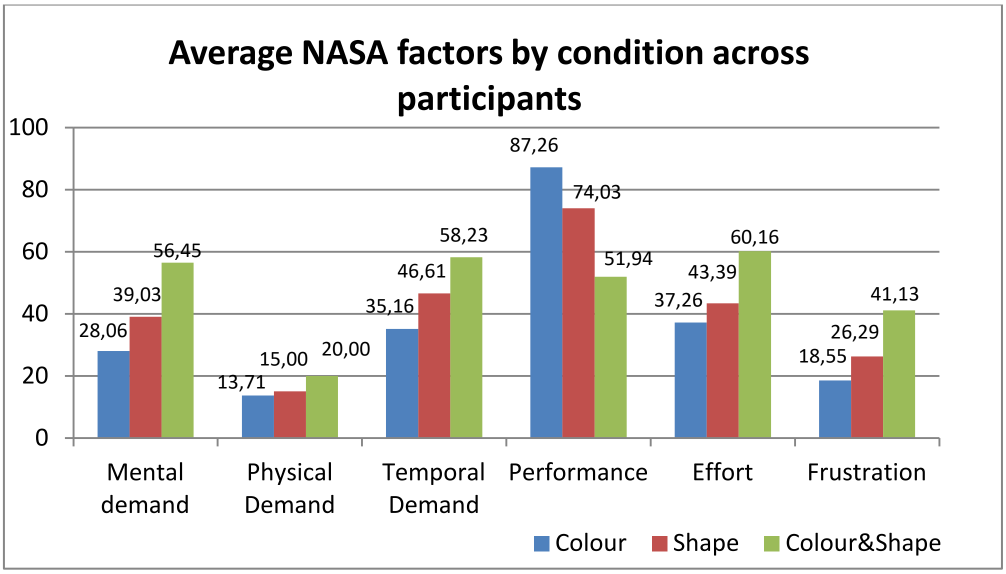


Figure 29. Average NASA TLX subscales for conditions *colour, shape* and *colour&shape* averaged across participants.

Observing the chart, condition *colour* is considered by all participants as the easiest task, resulting in all subscales to obtain the lowest values. The exception is the subscale "Performance" since this subscale answers to the question "How well did you perform the task?" Being condition *colour* considered as the easiest task, it is expected that it reflects the best Performance of all conditions.

On the contrary, condition *colour&shape* is considered the most difficult task, as it can be observed in the scores which reach the highest values of all three conditions. Likewise, Performance presents the lowest score, in line with the difficulty of the task.

Additionally, it is interesting to study if the subjective differences between the three conditions are statistically significant. In this regard, we have conducted a repeated measures ANOVA test, after testing our data for sphericity through a Maulchy's Sphericity Test. In order to detect where the differences are, we have also conducted a pairwise comparison with a Bonferroni correction. Results for the three tests can be seen in Table 23, 24 and 25 respectively.

Table 23 Results to test for sphericity in the data for our three conditions – *colour*, *shape* and *colour&shape*- and the measurement *NASA TLX.*

Measure: NASA TLX

Mauchly's Test of Sphericity

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Condition	0,882	3,130	2	0,209	0,895	0,956	0,500

Table 24. Results of a repeated measures ANOVA test for our three conditions – *colour*, *shape* and *colour&shape*– and the measurement *NASA TLX.*

Measure: NASA

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Condition	Sphericity Assumed	8465,311	2	4232,655	83,815	0,000	0,763
Error (condition)	Sphericity Assumed	2625,997	52	50,500			

Table 25. Results from the ANOVA test using a Bonferroni correction in order to find out where the differences between our conditions, for the measurement *NASA TLX,* occur.

Measure: NASA

Pairwise Comparisons

(I) Condition		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
Colour	Shape	-9,038	2,134	0,001	-14,499	-3,578
	C&S	-24,744	2,048	0,000	-29,984	-19,503
Shape	Colour	9,038	2,134	0,001	3,578	14,499
	C&S	-15,705	1,573	0,000	-19,730	-11,680
C&S	Colour	24,744	2,048	0,000	19,503	29,984
	Shape	15,705	1,573	0,000	11,680	19,730

The Maulchy's Test of Sphericity shows that the assumption of sphericity has not been violated (χ^2 (2) = 3.130, *p* = 0.209).

Furthermore, the results of the ANOVA test prove that there are statistically significant differences between our three conditions (F (2, 50.50) = 83.815; p = 0.000).

The pairwise comparison test helps us to find out where the differences appear. Observing Table 25, we can conclude that every condition differs significantly from each of the other two conditions. These statistical results are in line with the difficulty imposed by each task, differing each of the conditions in the cognitive load imposed by the difficulty of the task.

9. Discussion

In this chapter, we discuss, based on the results presented in Chapter 8, the applicability of each of the eye tracking measurements to assess the cognitive load imposed by our three conditions. We have found statistically significant differences between our three conditions and between OnRun/OffRun periods for each of our measurements that indicate the existence of a correlation between cognitive load and eye tracking measurements. However, the eye tracking measurement blinks partly contradicts our expectations.

In this chapter, we assess each research objective separately, discussing how each of the measurements can be applied to measure cognitive load. Moreover, based on the discussion of the individual research objectives, we address our overall research question.

Fixations

> Through the results obtained in the analysis of fixations we can answer our sub-question:
>
> *"How does the analysis of fixations help us to understand cognitive load?"*
>
> As we expected, there are significant differences between our easiest condition, searching by *colour* in comparison to searching by *colour&shape,* as the latter requires a higher level of cognition that is reflected in an increment in the number of fixations per second. This result is in line with the need to focus on two characteristics of the object (colour and

shape) instead of just in one of them, causing a higher workload. There are as well, statistical differences between condition *shape* and condition *colour&shape* that follow the same reasoning.

Furthermore, we have found statistical differences in the mean number of fixations per second for conditions *shape* and *colour&shape* and, in line with the results presented by Chen et al. (2011), we have found an increment in the number of fixations per second for periods of greater cognitive load, such as OnRun periods.

However, fixations alone cannot describe a difference in cognitive load between conditions *colour* and *shape* where there is just one characteristic to focus on when distinguishing the target from distractors. Although just fixations cannot describe overall changes in the level of cognition between those two conditions, it is interesting to continue investigating and find out if this measurement, in combination with other gaze events such as saccades or blinks, can be used to describe cognitive load in conditions where there is just one characteristic that determines the target item.

In the following chart, Figure 30, we can visually identify the difference between the mean number of fixations per second per task across our three conditions – *colour, shape* and *colour&shape.* In line with the results from our descriptive statistics presented in Table 6, condition *colour&shape* presents the highest number of fixations per second in relation to a higher cognitive state, while condition *colour* presents the smallest number of fixations per second, in relation to a lower cognitive state derived from the lower difficulty of the task.

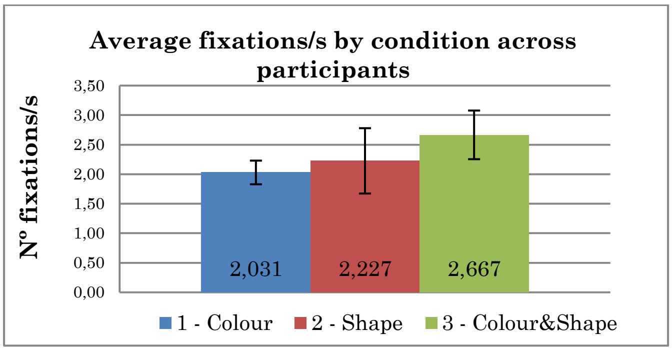


Figure 30. Summary of the number of fixations per second across participants for conditions *colour, shape* and *colour&shape.*

Saccades

"How does the analysis of saccades help us to understand cognitive load?"

From our analysis of saccades, we can conclude that the number of saccades per second can describe the level of cognition presented in a task.

Through this measurement, we can detect periods of higher cognitive load – OnRun periods – represented as a higher number of saccades per second, in comparison to periods that present a lower cognitive load, such as resting periods between runs in our tasks. We have found statistically significant differences between those periods for all conditions. Unlike for fixations, saccades already presented a significant difference towards the end of the task for our easiest condition, *colour*.

Moreover, there are significant differences in the number of saccades per second between conditions *colour* and *colour&shape* as well as between conditions *shape* and *colour&shape*. These results, in line with the argument for fixations, suggest that the difficulty level imposed by finding a target by two different characteristics – colour and shape – imply a higher level of cognition than searching for a target by only one characteristic – either colour or shape-.

A summary of the number of saccades per second for each of the conditions can be seen in Figure 31.

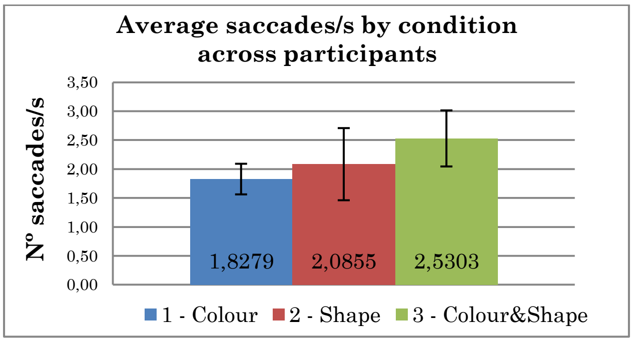


Figure 31. Summary of the number of saccades per second across participants for conditions *colour, shape* and *colour&shape.*

Despite the significant differences in saccades per second found in condition *colour* towards the end of the task, we can conclude that saccades alone cannot be used to distinguish between two tasks that offer a similar level of cognition (conditions *colour* and *shape)*. However, saccades can help us to distinguish between two conditions that offer a greater cognitive load, such as *shape* and *colour&shape.*

Additionally, we can relate our findings to the ones described by Rudmann et al. (2003), who claimed there exists a positive correlation between the number of saccades and cognitive load.

Blinks

“How does the analysis of blinks help us to understand cognitive load?”

The gaze event *blinks* offers similar results as *fixations* and *saccades*. However, one should be careful when describing cognitive load based only on the blink rate information.

Blinks are not adequate as a measurement to describe cognitive load if the duration of the task is not long enough, as we learned from our pre-study. In our experiment, we had an average task’s duration of 7 seconds, which is enough to detect statistically significant differences between short periods of higher level of cognition (OnRun) and low periods of level of cognition (OffRun).

For condition *colour* there are statistical differences in the blink rate from Run 8 on, as they were for *saccades*. Conditions *shape* and *colour&shape* offer statistical differences in the blinking rate for all runs.

Figure 32 presents an overview of the blinking rate per task for our three conditions.

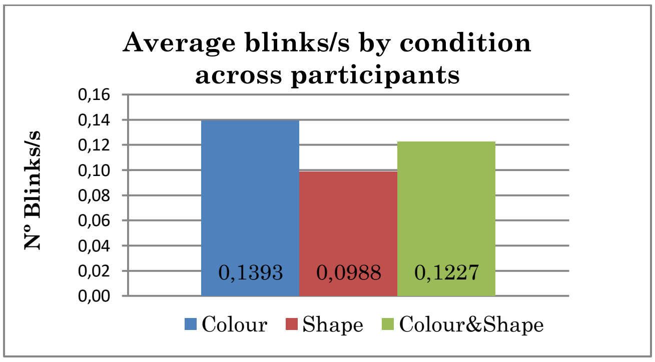


Figure 32. Summary of the number of blinks per second across participants for conditions *colour, shape* and *colour&shape.*

As one can appreciate in the chart based on the analysis of blinks, condition *shape* is cognitively more demanding than condition *colour&shape*, because fewer blinks are an indicator of a higher cognitive load. These results are not in line with the other objective measurements, *fixations, saccades, pupil dilation* and the subjective *NASA TLX*.

A possible explanation for these results could be imposed by the kind of task. In condition shape, participants cannot take the colour of the target as a dismissive characteristic and theredore, they need to focus more carefully. However, this finding is against our expectations, and since all our other objective eye tracking measurements and the subjective measurements gathered through the NASATLX questionnaires indicate condition *colour&shape* as the most challenging and demanding task, we would not consider blinks as a good indicator for cognitive load in such a visual search task.

However, it will be interesting to study this measurement in a different kind of task and its relation to cognitive load to find out if this condition is linked to visual search tasks or is generally applied. To our knowledge, there are no studies so far that relate the number of blinks to cognitive load, only to states of attention such as tiredness (García Barrios et al. (2004)).

Pupil dilation

"How does the analysis of pupil dilation help us to understand cognitive load?"

The pupil dilation has been demonstrated to be a great indicator of the level of cognition for short periods (On/OffRun). This fact is in line with our expectations since many studies identify pupil dilation as a valid indicator for cognitive load (Chen et al. (2011), Klingner et al. (2008), Pomplun and Sunkara (2003), Rafiqi et al. (2015), Rudmann et al. (2003)). It allows identifying how the pupil evolves together with the level of cognition, increasing in OnRun periods while participants search for the target item, reaching its maximum peak when participants find it and then decreasing in the resting period, as the level of cognition decreases too. Therefore, pupil size correlates well with changes in the level of cognition.

This measurement is not adequate, however, to describe the level of cognition at a task level, since pupil dilation is very sensible to changes in cognitive load and to changes in luminance, fluctuating therefore in time. It would not make sense to base the level of cognition in the average pupil diameter per entire task since we would average over periods of high and low cognitive load that would negate each other.

From the combined chart presented in the previous chapter, Figure 28 (see Appendix B), we can deduce that the pupil fluctuates in the same range for all conditions. However, one can find differences in its shape such as a smoother representation in condition *colour*, a peakier representation with greater differences between maximum and minimum values in condition *shape* and an increment in the overall fluctuation towards the end of the task for condition *colour&shape*.

If one wants to use the pupil diameter as an indicator to describe the level of cognition per task, it would be recommendable to average the dilation of the pupil based on the positive peaks reached by the pupil in its fluctuation, as these peaks correspond to maximum levels of cognition.

NASA TLX

How well correlate the analysis of fixations, saccades, pupil dilation and blinks to the subjective NASA TLX questionnaire?

As described in Chapter 8, we have found statistically significant differences between our three conditions, *colour, shape* and *colour&shape.* In line with our expectations when designing the task, condition *colour&shape* induces the highest cognitive load, and it is considered, as well, the most difficult task. Condition *colour*, on the contrary, is the task that induces the fewest cognitive load.

> The results obtained for the measurements *fixations, saccades* and *pupil dilation* correlate well with the results from the *NASA TLX* questionnaire. For all these measurements, we can appreciate an increment from condition *colour* to condition *colour&shape.*
>
> However, the results for the measurement *blinks* do not correlate with the results of the NASA TLX, in contrast to our expectations.

Based on the discussion of the individual research objectives, we address our overall research question:

"How can we measure cognitive load with eye tracking in visual search tasks?"

We can conclude that the cognitive load can be detected during periods of high mental workload (OnRun) such as performing a task to search for a target item, in comparison to periods of lower mental workload such as the period of time after the item is found (OffRun), when the participant is not actively searching anymore. We can base the detection of cognitive load in eye tracking measurements such as *fixations*, *saccades*, and *pupil dilation*, and correlate those measurements with subjective questionnaires such as *NASA TLX.*

Additionally, we consider *pupil dilation* a good indicator of cognitive load as a single measurement as well as the combined analysis of *fixations* and *saccades.* On the contrary, it would be necessary to study the measurement *blinks* in other scenarios, besides visual search, in order to relate it to cognitive load. As for now, it has not been related to cognitive load but only to states of attention.

A single eye tracking measurement may not be enough to describe the level of cognition itself, but what makes eye tracking powerful is the possibility to combine several gaze events and obtain results from all of them. For example, knowing that there is a level of cognition detectable when the number of saccades per second is equal or higher to (2.53 ± 0.48 sac/s) (such as in condition *colour&shape*) and combining this value with the number of fixations per second, we could get a great indicator to compare the level of cognition for this task among other tasks.

10. Conclusion and Future Work

Through the study of gaze events and the analysis of our empirical results presented in this master thesis, we can conclude that it is possible to measure cognitive load in visual search tasks through the analysis of eye tracking measurements, such as the number of *fixations*, the number of *saccades*, and the *pupillary response*. Such measurements describe well the level of cognition in interactive tasks when their analysis is combined.

Through our research, we have **learned** that the description of cognitive load works better for hierarchical tasks of short duration, such as searching for a target item (OnRun), clicking on it, and waiting for the next round to start (OffRun) than for longer duration, such as the entire task. Calculating the level of cognition at a task level implies defining an overall level of cognition for measurements that are not adequate for it, such as pupil dilation, as we discussed in the previous chapter.

With the results obtained, we can **contribute** to building better systems in HCI. Describing the overall cognitive load per task can be interesting to design better and more efficient interfaces. The combination of gaze events such as fixations and saccades can, quite accurately, describe the level of cognition for *the entire task*. Moreover, being able to catalog the level of cognition in short-term tasks, such as OnRun/OffRun periods, can be of great interest to, for example, detect periods of lower cognitive load when the user can be interrupted for notifications. In this regard, one could, for example, use our results to create an interruption manager.

In addition, the combined analysis of fixations, saccades, blinks and pupil dilation are a great indicator of the level of cognition *at a lap level*. These measurements reflect changes in cognitive load that the participant experience throughout the task and these changes are meaningfully different in task execution (On/OffRun). This meaningful difference achieved through the use of an eye tracker, cannot be otherwise achieved through overall measurements, such as task completion time or user ratings.

Through the conduction of our experiment, **new questions** have arisen: can our findings be extended through a deeper analysis of eye tracking measurements? For which other purposes can we use our results?

To the former question, we can argue that the findings from this study can be extended through a deeper analysis of our gaze events, such as fixations' duration, saccades' velocity, saccades' amplitude, blinks' latency, etc.

From the latter question, we came up with the idea that finding a correlation, not only between cognitive load and the number of fixations, saccades or blinks but also between the aforementioned gaze events, would offer the possibility to design a model, in which cognitive load can be more precisely described. Furthermore, with the design of a model to describe cognitive load through eye tracking, one could forecast a user's level of cognition in a non-intrusive manner, predicting periods of lower cognitive load in which the user could be disrupted without affecting the user's performance and emotional state.

In **conclusion**, we have found out that there is a link between cognition and eye movement control and that this information can be studied to detect cognitive states, being particularly useful to provide any system that attempts to facilitate HCI with more information about the user's cognitive activities.

References

Brünken, R., Plass, J., Leutner, D. (2003). *Direct Measurement of Cognitive Load in Multimedia Learning*. Educational Psychologist, 38 (1), 53–61.

Chen, S., Epps, J., Ruiz, N., Chen, F. (2011). *Eye activity as a measure of human mental effort in HCI.* IUI '11 Proceedings of the 15th international conference on intelligent user interfaces. Palo Alto, CA; United States.

Dawson, N. (2015). *Eye Tracking: What is it for and when to use it.* Usability Geek, May 7, 2015

Field, A. (2009). *Discovering statistics using SPSS.* Third edition. SAGE Publications Ltd.

García Barrios, V., Gütl, C., Preis, A.,Andrews, K., Pivec, M., Mödritscher, F., Trummer, C. (2004*). AdELE: A Framework for Adaptive E-Learning through Eye Tracking.* Proceedings of IKNOW, Graz, Austria.

Hart, S. G., and Staveland, L. E. (1988). *Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research.* Elsevier Science Publishers B. V. North-Holland

Hoaglin, D. C., and Iglewicz, B. (1987). *Fine tuning some resistant rules for outlier labeling*. Journal of American Statistical Association, 82, 1147-1149.

Holmqvist, K., Nyström, M. (2011). *Eye-Tracking: A comprehensive guide to methods and measures.* Oxford University Press, Oxford, UK.

Hornbæk, K. (2011) *Some Whys and Hows of Experiments in Human–Computer Interaction.* Foundations and Trends in Human–Computer Interaction Vol. 5, No. 4, 299–373.

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). *Task-evoked pupillary response to mental workload in human-computer interaction.* In Extended Abstracts on Human Factors in Computing Systems, CHI EA 2004 (pp. 1477-1480)

Klingner, J., Kumar, R., Hanrahan, P. (2008) *Measuring the Task-Evoked Pupillary Response with a Remote Eye Tracker.* Proceedings of the 2008 symposium on Eye tracking research & applications. ETRA 2008. Savannah, Georgia.

Miller GA, Galanter E, Pribram KH. *Plans and the structure of behavior.* New York: Holt, Rinehart and Winston, Inc; 1960

Pomplun, M., Sunkara, S. (2003) *Pupil Dilation as an Indicator of Cognitive Workload in Human-Computer Interaction.* International Conference on Human-Computer Interaction.

Porta, M., Ricotti, S., Perez, C.J. (2012). *Emotional e-learning through eye tracking.* Global Engineering Education Conference (EDUCON) IEEE, pp. 1-6

Rafiqi, S., Nair, S., Fernandez, E., Kim, J., Larson, E. (2015). *PupilWare: Towards Pervasive Cognitive Load Measurement using Commodity Devices.* Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments. *PETRA'15.* Island of Corfu, Greece.

Rudmann, D., McConkie, G., Zheng, X. (2003). *Eye tracking in Cognitive State Detection for HCI.* ICMI '03 International conference on Multimodal interfaces, Vancouver, British Columbia, Canada.

Sweller, J. (1999). *Instructional design in technical areas. ACER Press.* Camberwell, Australia

Toyama, T., Sonntag, D., Orlosky, J., Kiyokawa, K. (2015). *Attention Engagement and Cognitive State Analysis for Augmented Reality Text Display Functions*. Proceedings of the 20th International Conference on Intelligent User Interfaces *IUI 2015*, Atlanta, GA, USA

Appendix A

Welcome! – How to measure Cognitive Load with ET

Welcome to today's experiment. I am pleased with your attendance and I appreciate that you spend your time participating in this study. Your contribution to my study is essential and will help me to support my work. Before we get started, please read the following introduction to understand what this experiment is about and how it works.

Objective of the study:

The objective of the study is to understand how the presentation of the information influences your cognitive load and to measure it. We will make use of an Eye-Tracker (wearing glasses) that will record the movements of your eyes as well as the size of your pupil.

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 about yourself. Afterwards I will give you a brief introduction about the system that you are going to use (Eye-Tracking glasses and Perceptive-pixel monitor). Please, if you have any questions, don't hesitate to ask! The next step will be to perform the experiment.

First of all, we need to calibrate the eye-tracker. That means, making sure that the data obtained reflects what the position in the real world you are looking at. To do so, I will ask you to look at three specific points on the screen. Once the eye-tracker is calibrated, we can start with the experiment itself, where you have to find a specific figure among others under time-limited. It will consist of three different tasks. Each task has 10 runs and 3 laps per run. For each run, the number of elements shown on the screen will increase and there will be 3 tries (or laps) per run. The duration of the lap is fixed to 10 seconds and if you click on the right element, you will hear a beep sound. Between each lap, there will be a circle in the middle of the screen and you should place the mouse pointer inside it.

The task will consist of finding an element by a specific colour and shape, finding an element by a specific colour and finding an element by a specific shape. After each task, I will ask you to fill a short NASA TLX questionnaire.

At the end of the experiment, you will have to fill one last questionnaire about the use of the system and I will kindly answer any questions you might have, as well as listen to any remarkable comments that you want to point.

Time frame and compensation:

Completing the whole experiment has an approximate duration of 60 minutes. If you feel uncomfortable or dizzy, you can cancel the experiment at any point in time. Please just notify the study advisor.

After the completion of the experiment, you will receive a compensation for your help of 8 Euros per 60 minutes.

Finally, I thank you for your participation and wish you lots of fun!

Declaration of consent

ID:______

Information to study management:

Study advisor: Elena Barreras

Institution: Human Computer Interaction, University of Konstanz

Study procedure:

I would kindly point your attention on the subsequent study procedures: 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 questions 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 you any questions.

Declaration:

I was briefed 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 exclusively be used for aforementioned purposes and - with my consent – for internal presentations.

I hereby declare my approval with the above mentioned points:

________________________ ________________________ ________________________

(Name) (Date,place)

(Signature)

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

Elena Barreras /11/2016, Konstanz

(Name) (Date,place)

(Signature)

ID:______

Demographic questionnaire – Measuring Cognitive Load with ET

Personal Information:

Gender

- o Male
- o Female

Age

______ Years old

Height

______ meters

Profession

- o Bachelor Student
  Field: ____________________
- o Master Student
  Field: ____________________
- o Employee
- o Other: ____________________

Highest graduation level

- o Hauptschulabschluss
- o Mittlere Reife
- o Fachhochschulreife
- o Abitur
- o Bachelor degree
- o Master degree
- o PhD

Glasses / Contact Lenses

- ○ Glasses
- ○ Contact Lenses
- ○ None

If you have corrective lenses, please indicate your graduation

- ○ Left Eye: _________
- ○ Right Eye: _________

Are you colour-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
- ○ Several times a year
- ○ Never

ID:______

Post- questionnaire – Measuring Cognitive Load with ET

Which task did you find the most difficult?

- Finding by color & shape (find blue circle among squares, triangles & circles in different colours)
- Finding by color (find blue item among squares, triangles & circles in different colours)
- Finding by shape (find a circle among squares & triangles)

Why? Give at least one reason to justify your answer:

Which task did you find the easiest?

- Finding by color & shape (find blue square among squares, triangles & circles in different colours)
- Finding by color (find blue circle among squares, triangles & circles in different colours)
- Finding by shape (find red square among squares & triangles)

Why? Give at least one reason to justify your answer:

Did you feel pressured due to the time limit?

Why? Give at least one reason to justify your answer:

Which search strategy did you follow to find the target for each of the tasks?

Give at least one reason to justify your answer:

- Finding by color & shape (find blue square among squares, triangles & circles in different colours)

- Finding by color (find blue circle among squares, triangles & circles in different colours)

- Finding by shape (find red square among squares & triangles)

I felt comfortable wearing the Eye-Tracker glasses

Strongly disagree | ○ ○ ○ ○ ○ | Strongly agree

The using of the Eye-Tracker glasses have influenced my performance.

Strongly disagree | ○ ○ ○ ○ ○ | Strongly agree

The presence of the study advisor has influenced my performance.

Strongly disagree | ○ ○ ○ ○ ○ | Strongly agree

Are there any general comments about the study that you would like to mention?

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 were you?

Very Low Very High

Appendix B

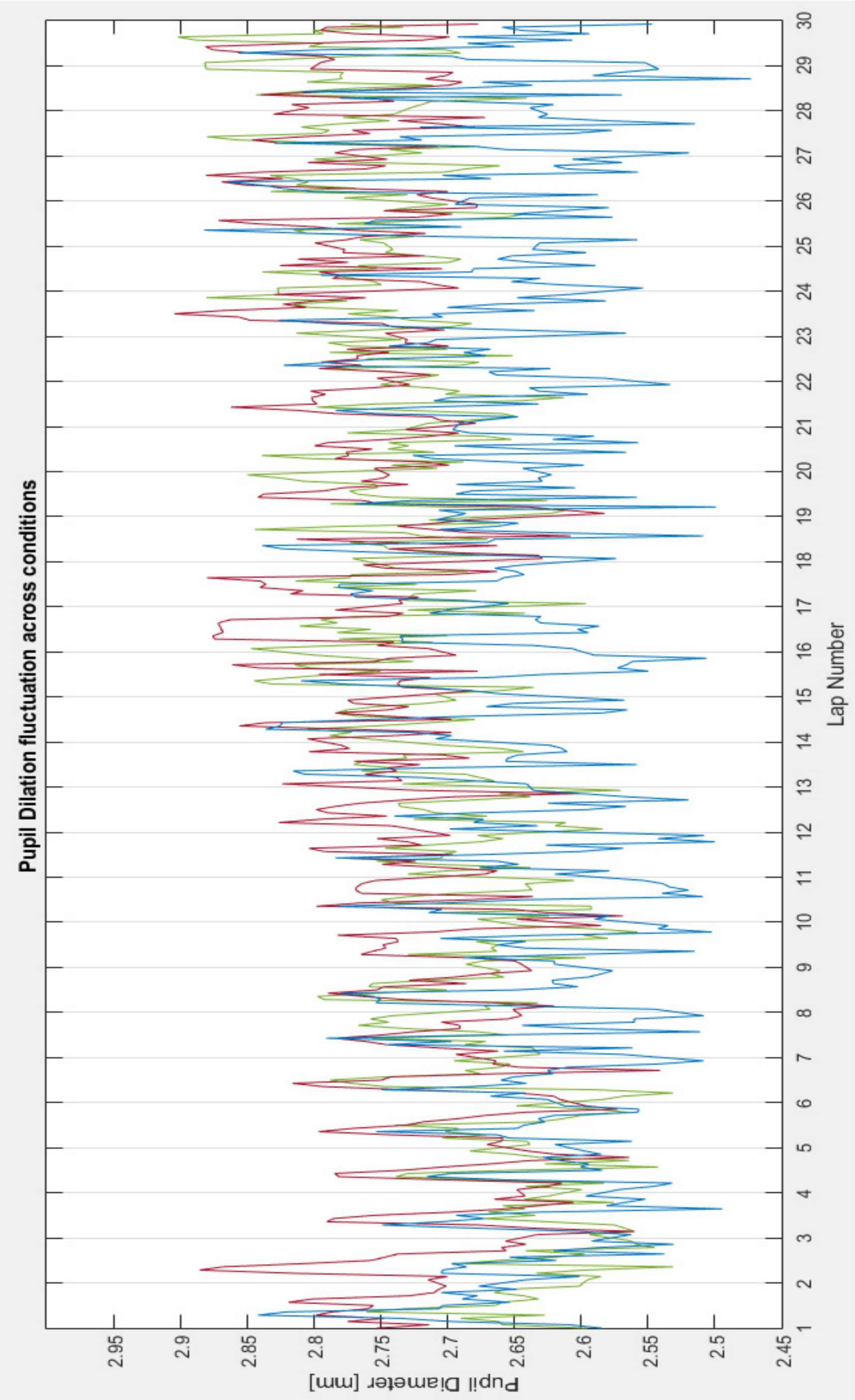


Figure 27. Average pupil dilatation for the conditions *colour, shape* and *olour&shape.*

Appendix C

Content of the USB flash drive

The USB flash drive contains the following folders and files:

- Thesis: the thesis document as a PDF file.
- Task: programmed task
- Logs: log files of the task
- ET Data: Exported Eye Tracking data
- SPSS: all files to be used in SPSS software for the statistical analysis.
- Knime: data files for the data formatting with KNIME
- Matlab: matlab scripts to automize the analysis
- Results: Exel files with the results of the questionnaires
- Study Documents: All documents using during the study