NTNU

Norwegian University of
Science and Technology

Estimating Mental Workload of University Students using Eye Parameters

Viveka De Alwis Edirisinghe

Master in Interaction Design
Submission date: June 2017
Supervisor: Frode Volden, ID

Norwegian University of Science and Technology
Department of Design

Preface

This thesis is submitted as a partial fulfilment of the master programme in Interaction design at NTNU Gjøvik. This was conducted during the spring semester in 2017. The original idea about studying the behavior of eye parameters with regards to stress, originated from the supervisor of this thesis, prof. Frode Volden, during the autumn semester studies in 2016. Therefore, as a foundation, I conducted an overall literature review regarding mental workload/fatigue in the IMT-4882:Specialization Course. This literature review lead me to choose the area path of studying further on doing an experimental research to estimate the mental workload of university students using eye parameters.

This report can be read by those who are interested in the field of mental workload and eye parameters, and curious about the new findings in the mentioned area. However, readers are assumed to have a basic background knowledge in statistical theories and eye parameters.

31-05-2017

Acknowledgment

I would like to thank and pay my sincere gratitude to my supervisor prof. Frode Volden for his kind guidance throughout the thesis, being flexible in arranging meetings in Oslo, to the support provided me to get lab accessibility to conduct the experiment, encouraging me in difficult situations, and all the other support given to me from the beginning to the end. I feel really lucky that I got the opportunity to work with him. Also, special thanks go to my friends who helped me during the good and bad times, making me up when I was in difficult situations. At last but not least, I would also like to pay my gratitude to all the voluntary participants who willingly participated in the experiment.

V.R. DE A.E.

Abstract

Estimating mental workload using eye parameters in different fields has become a significant study focus in the area of research. It is vital to discover the most reliable eye parameter/parameters that can be used to estimate mental workload. In this study, N-back tasks with four difficulty levels were designed to induce mental workload for a sample of 21 university students at NTNU Gjøvik. 17 eye parameters were measured using *SMI RED250mobile* Eye Tracker at a sampling rate of 250 Hz. Analyzed data indicate that peak fixation duration is the most suitable eye parameter to estimate mental workload. It has a negative relationship with the mental workload, where higher peak fixation duration can be observed at lower mental workload and lower peak fixation duration at higher mental workload. Moreover, blink frequency, blink count, peak blink duration, and pupil diameter show a significant positive relationship to the mental workload. Most of the saccade parameters failed to show a significant relationship, while fixation frequency, fixation duration, fixation count, blink duration, saccade velocity, and peak saccade amplitude showed a partial relationship with the mental workload.

Keywords: *eye parameters, mental workload, eye tracker, NASA-TLX, n-back*

Contents

List of Figures

List of Tables

1 Introduction

The high mental workload can lead to physical, psychological and social issues. Performing demanding tasks for a long period can cause stress and fatigue[1]. This can be a source for both health and performance issues on people. Poor performance can bring severe consequences for critical jobs such as driving, aviation, and surgical operations. For example, in the driving context higher mental workload and poor performance can be a cause of accidents[2], and in schools, it can affect results of students in examinations.

The relationship between mental workload and various physical measurements of the body has been researched using different user groups. Physical measures, such as heart rate, Electroencephalogram (EEG), respiration rate, alertness monitoring, skin conductance level have been explored to find a relationship with mental workload[3]. In particular, heart rate has been used widely in estimating mental workload in the driving context[3]. The majority of these studies were conducted using drivers as the participant group. In addition to that, different types of other user groups, such as students, pilots, military groups, and surgeons have been considered as participant groups for many studies.

Eye parameters is another physiological measure that has been used in the context of mental workload estimation. The most significant feature about eye parameter measurement compared to the other physiological measures mentioned above, is the possibility to use non-wearable eye trackers. This enables the researcher to measure the natural behavior of mental workload of users. Most of the studies conducted to find the relationship between eye parameters and mental workload used drivers as the target group. In addition to that, they have used pilots, surgeons, cyclists, and students.

However, the number of studies carried out on students as a target group is considered to be few compared to other user groups. Moreover, a broader comparison of different eye parameters and how they can be used to find the relationship with the mental workload of students is required. In doing so, it enables researchers to compare and determine a suitable eye parameter for mental workload estimation. Even more, it introduces new parameters that are useful for estimating mental workload. For example, peak values of the eye parameters such as peak fixation duration, peak blink duration, and peak saccade duration can be studied to check their ability to estimate mental workload. Thus, the purpose of this study is to find out the relationship between different eye parameters and mental workload of stu-

dents and hence report the most suitable eye parameter(s) for mental workload estimation.

1.1 Background and Motivation

The original idea about focusing on the eye parameters and stress came from the supervisor of the thesis Frode Volden, and as a background for the thesis, I conducted a literature review that focused on eye parameters with regards to mental workload/fatigue as a course fulfillment IMT 4882 in autumn 2016. Then, based on the knowledge acquired from the study, the thesis topic was narrowed down to estimating mental workload of university students using eye parameters.

Several areas have not been focused much in the literature with regards to estimating mental workload using eye parameters. First, a wide range of eye parameters has not been compared against the mental workload. Second, it is harder to find studies that have focused on the peak values of some eye parameters. In addition to that, few studies have used students as the user group for estimating mental workload using eye parameters. Moreover, it is interesting to find the validity of the previous studies done on the field. These reasons motivated me to conduct this study and research further on finding the relationship between eye parameters and mental workload of university students.

1.2 Purpose of the Study

The purpose of this study is to measure different eye parameters of a selected set of students at NTNU Gjøvik while performing an n-back task to find out the eye parameter which has the most significant relationship with the mental workload. Many studies have focused on discovering the relationship between mental workload and eye parameters. However, there is a need to explore a broad range of parameters and compare them against the mental workload. This will introduce new eye parameters that have not been discussed before, and provide extensive support for mental workload estimate in future.

To find the relationship between 17 different eye parameters and mental workload, I conducted an experiment which involves four different n-back tasks. The primary focus of these four various n-back tasks is to increase the workload of the user systematically. Twenty-one students selected from NTNU Gjøvik participated in the experiment. Each student's eye parameters at each n-back task level were recorded using a remote eye tracker, and they reported their mental workload by filling up a National Aeronautics and Space Administration-Task Load Index(NASA-TLX) form at the end of each level.

By collecting each user's eye parameter values and NASA-TLX feedback, the relationship between mental workload and each eye parameter is expected to find.

There can be both positive, negative or no correlation between the two factors. Another goal of the study is to experiment with the peak values of the eye parameters which have not been discussed extensively in other studies.

1.3 Research Questions

The purpose of the study can be formulated to three main research questions.

1. Which eye parameter, has the most significant relationship with the mental workload of a student?
2. Have saccadic eye parameters a significant influence on mental workload compared to blinks, fixations, and pupil diameter?
3. Has "*eye parameter x*" any significant relationship with the mental workload?
   H_0 : There is no significant difference in the "*eye parameter x*" between students who experience *low mental workload* and *high mental workload*.
   H_1 : There is a significant difference in the "*eye parameter x*" between students who experience *low mental workload* and *high mental workload*.

   Here, "*eye parameter x*" refers to the following eye parameters.

   - Blink count
   - Blink frequency
   - Blink duration
   - Peak blink duration
   - Fixation count
   - Fixation frequency
   - Fixation duration
   - Peak fixation duration
   - Saccade count
   - Saccade frequency
   - Saccade duration
   - Peak saccade duration
   - Saccade amplitude
   - Peak saccade amplitude
   - Saccade velocity
   - Peak saccade velocity
   - Pupil diameter

1.4 Contributions

The most useful eye parameter to estimate the mental workload more accurately than the other eye parameters will be discovered using 17 different eye parameters.

A wide comparison like this has not been done previously. Among 17 parameters, those with significant relationships to mental workload will be identified. In addition to that, already explored relationships in the literature will be checked for their validity. Moreover, new parameters which are not studied in previous research will be identified. The findings of the study can be used to design the interfaces of online learning systems, information systems, different types of websites in online shopping, and news sites more interactively without making users mentally overloaded.

1.5 Thesis Outline

The thesis consists of 6 chapters.

Chapter 1 Introduction discusses the motivation and background for the thesis, purpose of the study, research questions that are solved in the thesis, and the contributions of this thesis to the research field.

Chapter 2 - Background illustrates the previously conducted research based on three areas: mental workload, eye parameters, and eye trackers. These will be discussed in detail with related to the study of the thesis.

Chapter 3 - Methodology presents the experiment and analysis methods that the researcher has conducted to find the answers to the previously mentioned research questions.

Chapter 4 - Results reports the results of the data collected through the experiment. An extensive statistical data analysis is conducted for the collected data using descriptive statistics and inferential statistics.

Chapter 5 - Discussion presents the summary of the results and whether those findings reflect the validity of previous research or if there are any contradictions. Newly discovered findings are also presented.

Chapter 6 - Conclusion summarizes the final conclusions taken from the study and the data analysis. Research questions outlined in `Chapter 1` will be answered in this chapter. In addition to that, it further lists out possible future work.

2 Background

This chapter provides background literature of the study under three section. The first section presents the general topics on mental workload, such as its definition, different causes for mental workload and how to measure and generate mental workload. Next, various eye parameters measured are discussed illustrating various studies that have focused on those parameters. Finally, there is a detail description of various eye trackers and their specifications.

2.1 Mental Workload

Before discussing various studies on mental workload, it is important to understand the basic background of mental workload such as its definition and other terms referring the same concept.

2.1.1 Terms and Definitions

In literature, different terms have been used to refer to mental workload. One of the most commonly used terms was cognitive load[4, 5] or cognitive workload[6]. In addition to that, terms such as workload[7, 8], cognitive effort[9], and mental effort[10] have been utilized for the mental workload. However, the term mental workload can be considered as one of the most preferred terms[11, 12, 13, 14, 15]. Although there are various terms for mental workload, they refer to the same phenomena.

Similarly, there are various definitions for mental workload since it depends on the context of the usage. Pass & Van [16] define it as the load imposed on a person's cognitive system when a person performs a particular task. According to them, the cognitive load can be represented under three dimensions: *mental load*, *mental effort*, and *performance*. Mental load originates from the interaction between the learner and the task, while the mental effort is representing the actual mental capacity used when performing the task. Performance is the learner's achievement at the end of the completed task. Of these three dimensions, mental effort is the most reflective dimension of the mental workload as it involves the user's real mental effort allocated for the task. Performance can be affected by different mental effort. Therefore, some definitions of mental workload have defined it as an intervening factor of performance. For example, Parasuraman & Caggiano [17, as cited in 18] define mental workload as a state or set of states of the brain that intervene the performance of "perceptual, cognitive and motor tasks"[18, p.336]. Further-

more, Tokuda et al. [11] have defined mental workload as a concept that indicates the mental or cognitive busyness of a given person. In other words, the mental workload is the effort or perceived effort put to solve a problem/task by learning, thinking or reasoning[19].

2.1.2 Causes of Mental Workload

Different factors can cause higher mental workload. Task difficulty, time pressure, performance, age, physical effort, frustration, tension, fatigue and the type of activities can be considered as some of them[20]. Especially, time pressure and mental pressure can affect mostly for the mental workload. Bodala et al.[19] state that the pressure occurred during the execution of a memory task can result in generating mental workload. In such memory tasks, the participant's memory can be restricted due to the pressure. Chen & Epps[4] state that, in general, restricted working memory can lead to mental workload and generates while performing demanding tasks according to different user characteristics.

2.1.3 Implications of Mental Workload

The high mental workload can lead to physical, psychological and social issues. Jorna[1] emphasizes that performing mentally demanding tasks for an extended period can cause stress and fatigue. As a result, severe health problems such as hypertension and cardiac failures can occur. On the other hand, higher mental workload affects the performance of a person. A performance degradation can be severe for critical jobs such as driving, aviation, and surgical operations. Especially, in the driving context higher mental workload and poor performance can be a cause of accidents[2]. For students, poor performance in exams due to mental workload can be frustrating. Furthermore, the poor performance of doctors performing surgical operations can be life-threatening for patients. Therefore, identifying and measuring of the mental workload at the right time is significant.

2.1.4 Generating Mental Workload in Experiments

Before looking into different methods of measuring and estimating mental workload, it is worthwhile to study on various techniques that are suitable and not suitable to induce mental workload. At once, one might think that giving a difficult task to solve as a way of inducing mental workload. However, in general, such tasks could be solved by people who have a greater knowledge level on the subject. Therefore, they are not suitable for inducing workload, especially, for students. As an alternative, N-back tasks which were discovered in 1958 by Wayne Kirchner [21] can be used, and they are quite open and specially made for inducing mental workload. Jaeggi et al.[21] consider N-back task as an excellent indicator for working memory experimental research which works well in a higher level of men-

tal load. Within the memory capacity of a user, it is possible to let them use more mental effort on a task by increasing the complexity level[22]. Moreover, tasks that involve increasing time pressure, and tasks that combine several techniques have been used in different studies to generate mental workload.

N-back task can be designed in two different ways: (a) as an auditory n-back task or (b) as a visual n-back task. An auditory N-back task let the users hear different sounds repeatedly, and then they have to remember and recall the previous sounds. This method was used to control the mental workload by Tokuda et al.[11] in their experiment performed using 16 college students to find the correlation between saccadic intrusion and mental workload. They used up to four levels of N-back tasks where the last task(4-back) was used to generate the highest mental workload. Gable et al.[3] used an n-back digit recalling memory task(an auditory task) while driving to investigate the behavior of heart rate (HR) and pupil size, using a previously studied data of 8 students with a mean age of 21.1 years and with an average driving experience of 4.9 years. In this experiment, they used three n-back tasks: 0-back, 1-back, and 2-back to introduce the cognitive load which was successfully generated during the experiment.

Some studies have used a combination of different techniques to produce mental workload. This technique might have the capability to produce relatively higher mental workload as it involves extra difficulty and concentration. Thirteen volunteers having a mean age of 32.5 years(SD = 10.6years) participated in the experiment done by Tsai et al.[6] were supposed to perform a driving task, auditory task and both driving and auditory task to induce the mental workload. In the driving task, users were expected to maintain a specific distance between the front vehicle and the vehicle behind by driving at a constant speed. If the user could not keep the required speed, say for example they drive too fast or too slow, then the first vehicle crashes due to high speed and the vehicle behind crashes due to low speed. In the auditory task, they used a version of Paced Auditory Serial Addition Test (PASAT), where users were supposed to add the series of numbers played by the speakers and tell the answers loudly. In the next stage, users had to perform both driving and the auditory task. This combined experiment was an interesting, unique method that was used to introduce mental workload for the users.

Several other techniques, such as running, and imposing time pressure, have been used by different studies to generate mental workload. With the participation of 16 motorcyclists, Ohtsuka, et al.[15] conducted an experiment using two objectives to induce low mental workload and high mental workload. The high mental workload was achieved by a fast run, where the low mental workload was achieved by a solid run under a sufficient safety level at the Japan cycle sport. Each participant had to ride four laps (2 conditions * 2 runs). However, these types of

experiments can be considered as more physical than mental. For example, it might not be suitable for estimating mental workload of students. Comparatively, a time pressure task could be more appropriate in such case. For example, He et al.[23] have used time pressure when designing their experiment to introduce the mental workload for ten university students(min. age=19 years, max. age =25 years).

2.1.5 Techniques Used to Measure/Estimate Mental Workload

In general, it is not always easy to identify the amount of workload in a given situation for a specific person since the mental workload level is different from person to person based on their characteristics. However, it is possible to find out the relative level of mental workload, for example, whether a person experiences high mental workload or low mental workload. In essence, the mental workload can be estimated by measuring factors such as mental effort and performance[16].

One of the techniques used for this purpose is NASA-TLX forms. NASA-TLX form developed by Hart & Staveland[24] indicates six questions categorized under *Mental Demand*, *Physical Demand*, *Temporal Demand*, *Performance*, *Effort*, and *Frustration*. However, in a purely mentally demanding task, it might not be necessary to have the question related to *Physical Demand*. Each question can be answered by providing a rating score that ranges from 1 to 21. Scale value 1 represents very low, and 21 represents very high. `Figure 1` illustrates a sample of the NASA-TLX form.

Pfleging et al.[26] state that considerable amount of research have been conducted to measure the mental workload of individuals with the usage of successful different types of techniques. In particular, most of the experimental studies have used the NASA-TLX form to collect information from the users regarding the workload they experienced after completing the related task. At the end of the whole experiment, Pfleging et al.[26] collected feedback about all the tasks using NASA-TLX forms. However, it has to emphasize that this method is not a proper technique to gather feedback as it is harder for users to recall their prior mental state several n-back tasks ago. Therefore, it would be better to take the feedback from users at the end of each n-back task rather than waiting til the end.

In addition to NASA-TLX forms, some other scale systems have been used in some studies to assess users' mental workload. SWAT scale and Cooper-Harper scale have been used with regards to measuring mental workload [27, as cited in 23].

In addition to these scale systems, different studies have focused on different physical measurements of the participants to measure/estimate workload. Especially physiological measures such as heart rate, Electroencephalogram(EEG), skin conductance, respiration rate, alertness monitoring have been used to assess men-

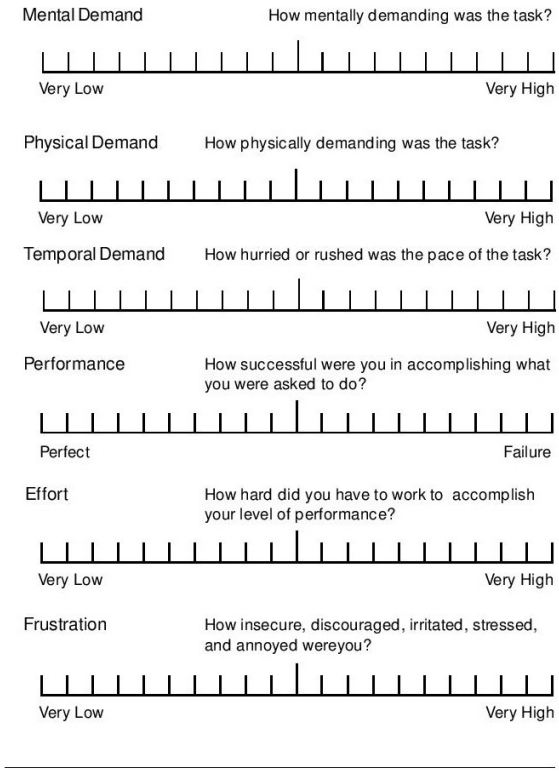

Figure 1: A sample of the NASA-TLX form[Source: NASA [25]].

tal workload[3]. In particular, most of these studies have been carried out using drivers and in the driving context. The majority of the studies have focused on heart rate as the dependent variable while the mental workload is the predictor[3]. However, the problem with these methods is they involve wearing various wearable devices on the participant to get accurate measurements. This might create uncomfortable situations among the participants which can ultimately lead to abnormal results.

On the other hand, eye parameters have also been used to estimate mental workload. The advantage of eye parameter measurement is, it does not necessarily need to wear physical devices. Measurement can be taken using a remote eye tracker or a camera. This enables the experimenter to take accurate results that correspond to the actual mental workload.

2.2 Use of Eye Parameters to Measure Mental Workload

Eye parameters can be categorized into four: blinks, saccades, fixations, and other. Eye parameters belong to these categories have been used to estimate mental workload. These parameters, their definitions and various studies related to mental workload are addressed in the following subsections.

2.2.1 Blink Count

Holmqvist et al.[20] define the term *blink* as the phenomena where the eyelid starts moving down to close the full eyeball, at which point, the pupil diameter, and corneal reflection cannot be measured anymore. Therefore, Holmqvist et al.[20] emphasize blink as a measurement of an eyelid movement rather than an eye movement. However, Bodala et al.[8] consider an eyelid movement as a blink when the pupil diameter is less than 0.5^0. Therefore, when a blink occurs, the pupil diameter value is either zero or closer to zero.

The relationship between a driver's mental workload and the behavior of the blinks have been researched in many studies. Blink is generally considered as a useful measurement of both fatigue, and mental workload[7].

In general, the measurement of blinks is either the blink count or blink frequency. The number of blinks that a particular user does while engaged in a task can be considered as the blink count. However, most studies have used blink frequency as the measurement for blinks.

2.2.2 Blink Frequency

The term blink frequency sometimes called as blink rate describes the number of blinks per second/minute[20]. Schneider & Deml[28] define blink frequency as the number of eye closures in a pre-defined period. Holmqvist et al.[20] summarizes previous research and concludes that the mean blink frequency of a typical reading

task lies around 3-7 blinks per minute and 15-30 blinks per minute for a non-reading task.

Blinks per minute were taken as the blink frequency by Zheng et al.[29] in their study carried out using 23 surgeons(M= 34.8 years, SD = 9.3 years). In the experiment, they used the median blink frequency(6 blinks per minute) to divide the data into two categories: low blink frequency group(<6 blinks per minute) and high blink frequency group(> 6 blinks per minute). They collected the feedback from the surgeons using NASA-TLX form, and the results showed that the group of surgeons who had low blink frequency, marked higher ratings in effort on performance, frustration level, and workload. Compared to the Holmqvist's blink frequency values mentioned above, the results show relatively low blink frequency(M= 6.4, SD = 5.8, min=0.2, max=25.8). However, blink frequency was not affected by the surgical task performance for both the groups. The reason behind this could be the high competence and experience(M = 3 years, SD = 3.6 years) of the surgeons for performing such tasks. On the other hand, in their research, they did not have a proper method of making the task difficult and see how the eye blinks behaved; rather they observed the blinks behavior within the given surgical procedure. Moreover, it should be noted that the method they used to split the data set using the median value is not considered as best practice. In other words, this technique of converting a continuous variable into a categorical variable using median value (median split) is not a best practice[30].

Tsai et al.[6] have found opposite results in their research conducted using 13 volunteers. As mentioned in section 2.1.4, they used a combined technique to generate mental workload among the participants. The results showed an increase in blink frequency while users were under higher mental workload conditions.

Ledger[31] discusses in his study carried out using 30 psychology students(mean age of 21.1 and 19.9) at Plymouth University, the importance of the results of blink rate with regards to cognitive workload. The other focus of his research was to determine the effect of audio information towards increasing mental workload as audio has been utilized in few researchers. Instead of using an eye tracker or a camera, Ledger has used Electrooculography(EOG) to measure eye blinks. In this method, several electrodes are connected to the eyes and ears to record the measurements. Results indicated that blink rate decreased with the increment of mental workload which contradicts the early findings of Tsai et al.[6]. Furthermore, he claims that these conclusions can be used as a communicating method for those who cannot communicate properly or those who cannot express what they feel. Moreover, he mentions that there have been very few researches done on finding the relationship between the mental workload and blink rate.

However, factors such as dry eyes, air pollutants, contact lenses, use of monitors,

time of task and time of day can affect the blink rate in addition to the mental workload[20]. Therefore, it is quite important to find the effect of mental workload alone towards the measured parameters.

2.2.3 Blink Duration

Holmqvist et al.[20] define blink duration as the time taken from the eyelid starts moving down to the point where it returns the original position. Blink duration is a better and more sensitive parameter to predict workload compared to blink frequency[7]. Benedetto et al.[7] claim that significant results can be found in higher short blinks in the presence of higher mental workload. In other words, mean blink duration tends to decrease with the increment of higher mental workload. Typically, participants tend to do fast blinks when they are performing a highly cognitively demanding task as they are afraid of losing information due to long blinks. This resulted in low mean blink duration at higher mental workload.

2.2.4 Fixation Count

Fixation is a type of eye movement that is used to observe a specific visible area, and it is further classified into four types by Tokuda et al.[11].

- Tremors
- Slow drifts
- Micro-saccades
- Saccadic Intrusion(SI)

Similar to blink count, fixation count is correlated with the fixation frequency. The number of studies conducted on both of these is very few. However, a study done by Wang et al.[32] using 42 students at University of Southern China, focused on finding the relationship between fixation count and mental workload. They expected higher fixation count under higher task difficulties but did not manage to find significant results to satisfy the alternative hypothesis.

2.2.5 Fixation Frequency

Fixation frequency is the number of fixations done per unit time. Fixation frequency increases when the mental workload is high and decreases when the user is overloaded with mental workload[23]. The overloaded situation can be considered as a special one in mental workload. In general, the main focus in mental workload study has to be on the gradual increase in mental workload.

2.2.6 Fixation Duration

Fixation duration is defined as the time duration the eye is staring at a specific point, and it is considered to be the most used eye measurement in the field of eye tracking research according to Holmqvist et al.[20]. Schulz et al.[33] discovered in

their study that fixation duration decreases with the increment of mental workload. However, in this study, they were expecting the opposite, and therefore this result was a contradiction.

2.2.7 Saccades

Saccades are also a type of eye movement that shifts the gaze fixation point from one location to the other[11]. According to Tokuda et al.[11], saccadic eye movements have been discussed under three topics.

- Micro-saccades
- Regular saccades
- Saccadic Intrusion(SI)

Bodala et al.[8] investigated the relationship between micro-saccades and mental workload using the data collected from a previous study. They define the term micro-saccades as "small jerky, involuntary eye movements ($\sim 1^0$) that occur during fixations at a rate of 1 or 2 per second" [8, p.7994].

2.2.8 Saccade Frequency

Saccade frequency can be defined as the number of saccades occurred during a unit time. He et al.[23] discovered that saccade frequency increased with higher time pressure (higher time pressure indicates higher mental workload), and it decreased once the user was overloaded.

2.2.9 Peak Saccade Amplitude

Cardona & Quevedo[34] tried to discover the impact of both cognitive load and large saccadic amplitude towards blink rate. A driving task which included five difficulty levels based on traffic intensity and driving difficulty was performed using 20 users who were non-commercial drivers. Though the results could not find any significant results what they were investigating, they figured out that the count of blink-saccade pairs and number of large amplitude saccades increased by task difficulty (i.e., when the workload is increasing).

2.2.10 Saccade Velocity

Di Stasi et al.[18] and Holmqvist et al.[20] discuss the studies that have discovered the relationship between saccadic velocity and many other factors. These factors are arousal level (sometimes referred as cognitive level), fatigue, Rapid Eye Movement (REM) sleep, task difficulty level, anticipation, drugs and alcohol, different disorders in the clinical domain, military, and everyday tasks. Some or most of these factors might affect mental workload. However, studies conducted to find the relationship of saccade velocity on mental workload are very few and not that much of concrete findings. Therefore, it is an open area to do further research.

2.2.11 Peak Saccade Velocity

Peak saccadic velocity is a good indicator of arousal, ergonomics and in the clinic[18]. Although lower peak saccadic velocity was noticed in higher mental workload situation, Di Stasi et al.'s [12] research failed to find any significant results of saccadic velocity and saccadic duration on mental workload.

Among the measured values of saccade count, saccadic amplitude, saccadic duration, peak saccadic velocity, fixation count, fixation duration, pupil diameter, Di Stasi et al.[13] discovered that peak saccadic velocity is a more desirable indicator of mental workload of individuals compared to other parameters. The purpose of their study was to investigate which type of eye parameters were associated with the mental workload and different types of risk behaviors using a riding simulator. The experiment conducted by Di Stasi et al.[35] discovered that peak saccadic velocity decreased with the increment of mental workload.

In contrast, Bodala et al.[19] have discovered in their experiment that peak saccade velocity increased with the mental workload. However, after the user got overloaded in the task, the value of saccadic velocity started to decrease[23].

2.2.12 Pupil Diameter

In the literature, the majority of the research have discussed the relationship between mental workload and the behavior of the pupil diameter/dilation/size, and it was expected to increase with the increment of the mental workload.

The opening of the iris is known as the pupil, and it controls the amount of light that enters the eye[36]. Therefore, lighting condition becomes a significant factor that matters a lot when measuring pupil diameter. Holmqvist et al.[20] claim that the difference arises in pupil diameter with regards to cognitive or emotional effects are very low, and most of the changes occur due to the change in light intensity. Gable et al.[3] claim that the reason for the distraction of the values of pupil size is the variation of the amount of light that comes to the eye which is referred as Pupillary Light Reflex (PLR). Therefore, it is vital to maintaining the same brightness (sometimes referred as illuminance and luminance level) during the experiment.

Dilation of pupil diameter (which is also referred to as Task- Evoked Pupillary Response (TEPR) by Beatty[37]), occurs when someone is subjected to high mental workload[5]. The experiment conducted by Gable et al.[3] using few participants in a driving task, discovered that pupil size increased when the mental workload was high and suggested that it is a good indicator of the workload in real-time driving conditions. However, Tokuda et al.[11] claim that driving under different lighting conditions affects the pupil size, therefore conclude that SI is a better parameter when estimating mental workload compared to pupil diameter since

SI is independent of lighting conditions.

Since lighting conditions do affect the changes in pupil diameter, Pfleging et al.[26] conducted an experiment where the behavior of pupil diameter was observed under six different lighting conditions. An auditory n-back (0-back, 1-back, 2-back) digit recalling task was used to induce the mental workload, and each user had to undergo 6 (lighting conditions) $\times$ 3 (difficulty levels) = 18 trials during the experiment. Results indicated that the size of the pupil diameter increased with the task difficulty and it applies to all different lighting conditions. The equation that is suggested by them helps to calculate the actual pupil dilation value that creates only due to the difficulty of the task.

A study to find out the behavior of pupil diameter with regards to the suturing proficiency level was conducted by Cao et al.[38]. They discovered that the participants with experience were less stress while performing the task compared to the participants with less experience. Time spent to complete the task was measured and used as an indication of the proficiency. Their findings revealed that the pupil diameter is a closely related parameter with regards to measuring the suturing proficiency of each.

However, it is not only the mental workload that affects the changes in pupil diameter, but some other factors such as drugs, age, pain, diabetes, drowsiness & fatigue, and emotion & anticipation also have effects on it [20]. Therefore, it is necessary to conduct any analysis based on the assumption of those factors mentioned above, so that they have minimum effect on the mental workload. Therefore, it is important to use right equipment during the measurement of eye parameters.

2.3 Usage of Eye Trackers

Eye trackers are used to measuring different eye parameters. There are three types of eye trackers: static eye trackers, head-mounted eye trackers, and remote eye trackers. The main advantage of using remote eye trackers instead of head-mounted eye tracker or static eye tracker is that they prevent the user's abnormal behavior and let the user stay in a natural position. However, there is an advantage of measuring physiological measures such as heart rate when compared to using eye trackers. These physiological measures are not influenced by lighting conditions, unlike pupil diameter[26].

Different type of eye trackers has been used in the previous research when measuring the relationship between mental workload and eye parameters. There are diverse kinds of eye trackers in different frequency levels, which come under the categories of tower mounted eye trackers, remote eye trackers, and head-mounted eye trackers. Based on the research model and possibility, the most relevant eye tracker has to be chosen. Some of the eye trackers that have been used in the

previous studies are discussed in the following.

Tokuda et al.[11] have used Tobii 1750 eye tracker (50Hz) to their experiment with a chin support to limit the head movements which might affect the results. This eye tracker is a non-intrusive eye tracker which is located in front of the participant performing the task. EyeLink 1000 eye tracker has been used by Bodala et al.[19] which includes chin and head support, eye tracker PC and a video camera. Schulz et al.[33] used an EyeSeeCam (300Hz) head-mounted mobile eye tracker in their study experiment with the participation of trainee anesthetists.

An experiment conducted by Zheng et al.[29] with the involvement of surgeons to measure their mental workload with regards to NASA-TLX and blinks was done with the support of head mounted Locarna PT-Mini eye tracker (30 Hz). Users were supposed to wear a pair of goggles while they were performing the suggested surgical task which was attached to the eye tracker. Two small cameras were included in the goggles where one recorded eye movements regarding blinks and the other recorded the scene observed by the user.

Benedetto et al.[7] have used *SMI iView X HED head mounted monocular eye tracker (200HZ)* to get the eye measurements while users were performing the driving task using a driving simulator. Fifteen participants with driving experience took part in the experiment where the intention to find out the effect of using in-vehicle information systems while driving and how eye blinks behave with the increment of the driver visual workload. However, Benedetto et al.[7] claim that to guarantee the findings that they discovered regarding the blink duration should be tested again using a remote eye tracker with a larger sample.

However Pfleging et al.[26] claim that remote eye tracker is most reliable when it measures the pupil and they used SMIRED250(120Hz) eye tracker where iView X software recorded the data. Not only measuring pupil diameter, but remote eye trackers also help to get more reliable data irrespective of changing the user's natural behavior. In that sense, SMIRED250 eye tracker can be considered as a suitable eye tracker for remote eye tracking. The importance of a remote eye tracker while conducting the experiment is they forget the fact that their eye parameters are measured, and thereby end up with more reliable data.

The experiment conducted by Di Stasi et al.[12] with the participation of 46 Granada University undergraduates used Eye Link II head mounted eye tracker (500Hz), with the 13 point calibration to measure the eye parameters. However, their analysis was not conducted on the fixations around blinks, fixations, and saccades (where saccade duration is $<$10ms and saccade duration $>$ 100ms). Di Stasi et al.[35] used the same type of eye tracker in air traffic controller simulated multitasks with regards to mental workload.

3 Methodology

There has been much research performed on finding the relationship between mental workload and eye parameters. These lack two things that need to be addressed further. First, peak values of parameters, such as peak fixation duration, peak saccade duration, peak blink duration are hardly studied in the literature. Second, most of these studies have not compared a large number of different eye parameters within one study. On the other hand, it is not simple to find the most significant eye parameter that is related to mental workload without comparing a wide range of eye parameters.

Therefore, in this study, the main focus was to collect a wide range of eye parameter data under various mental workload levels to search for the most significantly related candidate for measuring mental workload. Among others, this range of eye parameters consisted of peak parameter values such as peak saccade duration, peak fixation duration, and peak blink duration.

This study followed a quantitative experiment performed by a group of participants who repeatedly exposed to difficult and challenging mental workload situations. In the experiment, four levels of n-back tasks were used as a means of introducing workload on the participants, and different eye parameters were measured under each n-back task. At the end of each task, participants reported their mental workload by submitting the online NASA-TLX form presented on the website. The collected data were analyzed using both descriptive and inferential statistics.

3.1 Experiment Setting

The experiment took place in a separate laboratory room at NTNU Gjøvik. Inside the lab room, each selected student performed the experiment independently and without any distraction from the outside world. A special "DO NOT DISTURB" notice was put on the door entrance to avoid disturbances. The illumination of the light in the lab room and the screen luminance of the computer were kept constant for every user to prevent additional effects. The room had a table for placing the practice session laptop and the experiment laptop. In front of the practice session laptop, there was a swivel chair. A regular fixed chair was set in front of the experiment laptop to achieve minimum body movements, and hence minimize the effect for the measurements. Figure 2 illustrates how some selected participants performed the experiment.

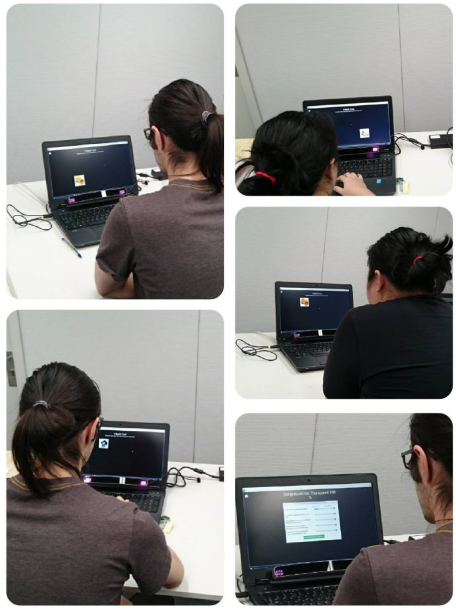

Figure 2: Experiment setting and several students performing the experiment.

3.2 Participant Selection

Participants for the study were selected using several ways. Some of them were colleagues of the researcher, and some were friends of researcher's friends (snowball sampling). The rest were selected using convenience sampling technique. Students who were at the university premises were selected randomly. Thirty minutes appointments were made with the participants, and each was given a premium in return for his or her voluntary participation.

Altogether, 21 students (12 males, mean age = 25 years, min = 23 years, max = 34 years, SD = 2 years) participated in the study. In general, these students were of various ethnic backgrounds. The majority of the users were from Norway, and the others represented Vietnam, Iceland, India, China, Myanmar, Kosovo, Macedonia, Mexico, Croatia, and Ukraine. Of the selected students, seven were having left dominant eye, and the rest were right dominant. Of the 21 students, nine had previous experience with this type of study as they participated in the pilot experiment performed before this experiment. All the participants had normal or corrected to normal vision.

3.3 Experiment

The independent variable manipulated in the experiment was each student's mental workload. This was achieved by the four different n-back tasks: *1-back*, *2-back*, *3-back*, and *4-back*. 1-back was supposed to have the lowest mental workload. 2-back task a bit higher workload, and 3-back task even higher workload than both 1 and 2-back tasks. Finally, the highest workload was expected from the 4-back task.

Each student's eye parameters were measured during each n-back task. These specific eye parameters were considered as the dependent variable in the study. Following are the eye parameters measured and analyzed during the study.

1. Blink count
2. Blink frequency
3. Blink duration
4. Peak blink duration
5. Fixation count
6. Fixation frequency
7. Fixation duration
8. Peak fixation duration
9. Saccade count
10. Saccade frequency
11. Saccade duration
12. Peak saccade duration
13. Saccade amplitude
14. Peak saccade amplitude
15. Saccade velocity
16. Peak saccade velocity
17. Pupil diameter

Typically, in an n-back task, users get to see a series of images. They have to remember each passing image, and if the current image is the same that was shown n times ago, then it will be a correct n-back match. For example, if your current image is the same as the previous image, then it is a 1-back match. If the current image is the same as the image before the previous image, then it is a 2-back match and so on.

Although there are already designed websites such as psytoolkit[1] that can perform the n-back tasks, a special website[39] was designed for this study. An example page of the designed website is shown in `Figure 3` (see Appendix A for a sample list of pages). The main purpose of not using the existing toolkits was their unreliability. For example, if something went wrong with the selected tool, there

[1] http://www.psytoolkit.org/experiment-library/nback.html

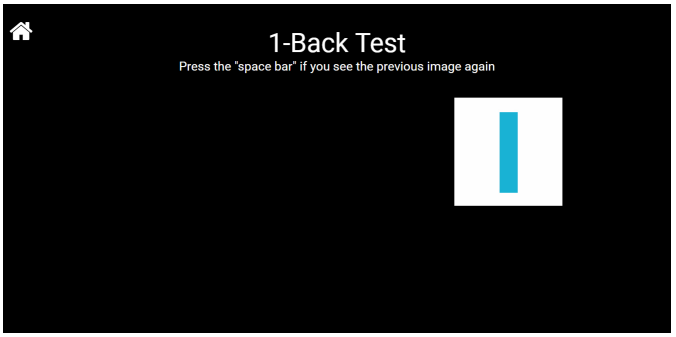

Figure 3: A sample page of the specially designed website to perform n-back tasks.

was no way of correcting that error. Rather it would be easy if we had our toolkit and to have the control over it.

Prior to the experiment, a pilot test was performed using 16 users with only three levels of n-back tasks. A series of 36 images consisting of various *fruit types (colored)* were used in the 1-back task, and *pictures of the alphabet (black and white)* and *vegetables (colored)* were used for 2-back and 3-back tasks respectively. Based on the experience and results obtained from the pilot test, following modifications were implemented to the formal experiment.

- Black and white alphabetic letters were changed into **colored letters**.
- **Images of fruits, vegetables, and letters were mixed** and designed the each level of the task using **40 images**.
- **The time duration between images** was reduced from 5 seconds to **4 seconds**.
- In each n-back task, **50%** of the images are **letters** and the **rest is a mix of fruits and vegetables**.
- **4-back task** was added to increase the workload
- The sample was restricted to **only students**, removing non-students.

Before conducting the experiment, every user was given an informed consent form (see Appendix B) for signing. Then the dominant eye of every user was found by making a small triangle using both hands and looking at a given stimulus through it and at the same time moving it towards the eyes.

Next, participants were given a short introduction on how the entire experiment is carried out and then performed a practice test[40] first on a separate computer. Images were shown randomly in one of the four locations specified on the screen as illustrated in `Figure 4`. However, a condition was set in such a way that the next image in the sequence will not appear in the same location as the current image.

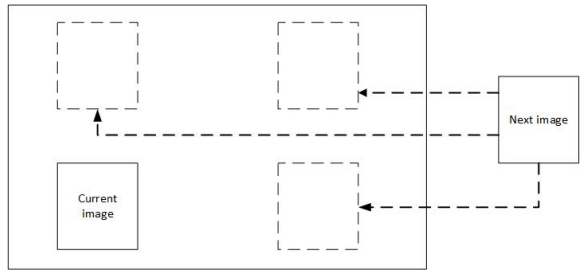

Figure 4: Images can appear in one of the four location. The next image cannot appear in the current image location but in one of the other three locations.

Participants could perform the practice session as many times as they want to get a full understanding of the task. The practice session was designed in the same way like the real experiment, but with using only a few images. Practice sessions were not recorded, and it was carried out on a separate computer.

In the experiment, participants were instructed to sit in a comfortable position. Then the distance between the user and the PC was kept around 60-70 cm for every user. They were asked to maintain the eye focus within the screen area, and if any problem occurs just ask the researcher while looking at the display so that she can clarify any issues. This worked well since results proved that users' focus within the experiment was over 95%.

Prior to the commencement of the real experiment, each participant had to undergo an eye calibration after their registration in the eye tracker PC. 5-point calibration method was used to do the calibration.

As stated above, the experiment of the study used a sequence of 40 images which were shown between 4 seconds interval, for all the four tasks. If the student saw the corresponding n-back image, they were supposed to press the space bar on the keyboard. Forty images consisted of 50% alphabetic characters, 25% representing fruits, and the rest as images of vegetables.

In the experiment, lighting conditions were kept constant for all users in the room. Users were asked to minimize unnecessary head movements. Otherwise, it could have affected the results if users shifted their focus from the screen. For instance, if someone knocked the door. However, a chin support was not used during the experiment due to the following reasons

- Letting users stay in an unnatural position by using a chin rest might affect their normal behavior, and it might make them extra physically stressful which is not a goal of the designed experiment.
- To avoid extra bias parameters that might affect the results

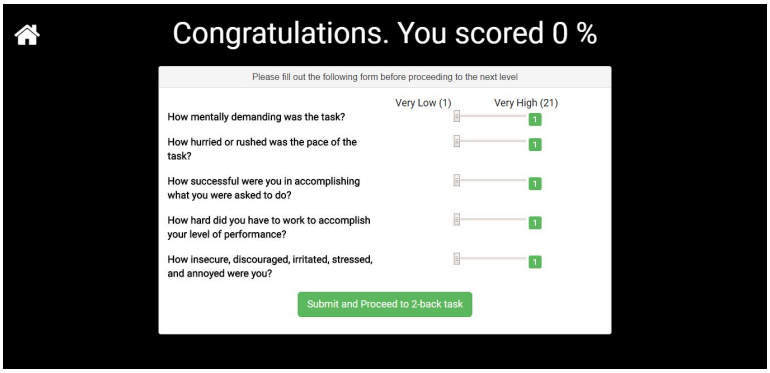

Figure 5: NASA-TLX feedback form which is shown at the end of each n-back task.

The designed experiment took not more than 15 minutes, and the total time taken for the recorded experiment was approximately 10 minutes. However, some participants took longer than usual while some took less. The main reason was that some participants spent a little bit of extra time for thinking while they were filling NASA-TLX form. Nevertheless, the time duration for each n-back task was equivalent for everyone.

3.4 NASA-TLX Form Ratings

At the end of each n-back task, users were prompted to fill in their ratings for five questions in the NASA-TLX form (see `Figure 5`). The rating was ranging from 1(Very Low) to 21(Very High). NASA-TLX forms are typically used to get an overview of the mental workload users experienced during a cognitively challenging task. Therefore, I have used the same technique to measure the mental workload in this study. Note that the standard question in a NASA-TLX form regarding the physical demand was not included in this form since the task does not require any physical effort.

3.5 Eye Parameter Measurements

The measuring of the dependent variables (eye parameters) was carried out using a remote eye tracker connected to the experiment laptop. In this study, SMI RED250mobile Eye Tracker was used, and the reasons for using it are elaborated in the following.

- This eye tracker is specially designed for studies conducting in the field of saccades.
- It is easy to use since it is portable.
- Eliminate the user's abnormal behavior by letting them sit and doing the task

in a natural way.

The sampling rate of the eye tracker is 250Hz and gaze position accuracy is 0.4^0[41]. The monocular eye-tracking mode was chosen as the mode for the experiment. This eye tracker comes with its associated laptop.

Experiment Center 3.6 software was used to perform the experiment, and the data analysis for the obtained data from the eye tracker was done by the BeGaze 3.6. Later on, these collected eye parameter data was exported to a format that can be used by the Statistical Package for the Social Sciences(SPSS) software for statistical analysis.

3.6 Data Analysis

The collected data was analyzed statistically using both descriptive and inferential statistics. The exported data from the eye tracker was imported to the SPSS package and analyzed. Because this study was a repeated-measures design within the same participants, paired-samples t-test and Analysis of Variance (ANOVA) for repeated measures together with Multivariate ANOVA(MANOVA) whenever necessary were used as the main inferential analysis techniques. Based on the results, rejection or accepting the null hypothesis was done by checking the significance of the mean difference of the eye parameters between different mental workload levels. ANOVA test results were used to find any significance between each pair of n-back levels. ANOVA test was not performed for those who did not have significant results in the paired-samples t-test. Screenshots of two samples of the carried out tests in SPSS are shown in Appendix C. However, several assumptions had to be fulfilled to perform these tests[30].

- Additivity and Linearity
- Removing of outliers
- Normality of the distribution of mean difference
- Homogeneity of variances
- Sphericity (only for ANOVA)

During the calculation of the significant probability, it is important to specify the effect size of the relationship. Therefore, the *Cohen's d value* was used to calculate the effect size for paired-samples t-test, and omega squared(ω^2) value was used for ANOVA. Field[30, p. 588] defines the equation 3.1 to calculate the ω^2 value.

$$\omega^2 = \frac{\left[\frac{k-1}{nk}(MS_M - MS_R)\right]}{MS_R + \frac{MS_B - MS_R}{k} + \left[\frac{k-1}{nk}(MS_M - MS_R)\right]} \tag{3.1}$$

where MS_M represents the mean square for the model, and the residual mean

square by MS_R. Sample size is n, and k is the number of conditions in the experiment. MS_B is the mean square between participants.

4 Results

Data collected from the experiment were analyzed using both descriptive and inferential statistics. In each section of this chapter, results of the repeated-measures ANOVA test for the four n-back tasks are presented first. If there was no statistically significant result, then the results of the multivariate test are discussed if appropriate. In addition to that, the results of the paired-samples t-tests done for the low mental workload and high mental workload groups are covered in each section. Before analyzing the results of each eye parameter, NASA-TLX form ratings were tested statistically to get a logical grouping of the mental workload.

4.1 Student Ratings for NASA-TLX Form

NASA-TLX form consisted of totally six questions, and only five of them were used in the experiment. Four of these five questions were used as indicators to measure each student's mental workload. Question 3 : "How successful were you in accomplishing what you were asked to do?", was not used as a measurement of the mental workload because it reflected the user's self-evaluation on how successful he/she was on accomplishing good performance.

4.1.1 Question 1: How mentally demanding was the task?

This question asked students about their mental workload or mental demand they experienced during the experiment. Descriptive statistics were calculated for all four levels of the n-back tasks to get a general idea of the ratings given to this question. One record was removed from the dataset as it was an outlier. The mean ratings and standard deviations were as follows (N=20): 1-back (M = 3.55, SD = 2.82) , 2-back (M = 7.00, SD = 3.68), 3-back (M = 10.90, SD = 4.24), and 4-back (M = 16.40, SD = 4.35). On average, mental demand was highest for the 4-back task and lowest for the 1-back task. According to the ANOVA test results, there was a significant difference between the four n-back levels regarding the mental demand ratings, $F(1.97, 37.35) = 88.22$, $p < .001$, $\omega^2 = 2.53$. Students experienced significantly demanding mental load in 2-back task compared to the 1-back task($F(1,19) = 34.54$, $p < .001$), and 3-back task had even more mental demand on them than 2-back($F(1,19) = 41.34$, $p < .001$), and even more in the 4-back task, ($F(1,19) = 55.53$, $p < .001$).

Moreover, the combined mean ratings of 1-back and 2-back($R_{1,2}$) and 3-back and 4-back($R_{3,4}$) for the sample of 21 students were calculated to find any signif-

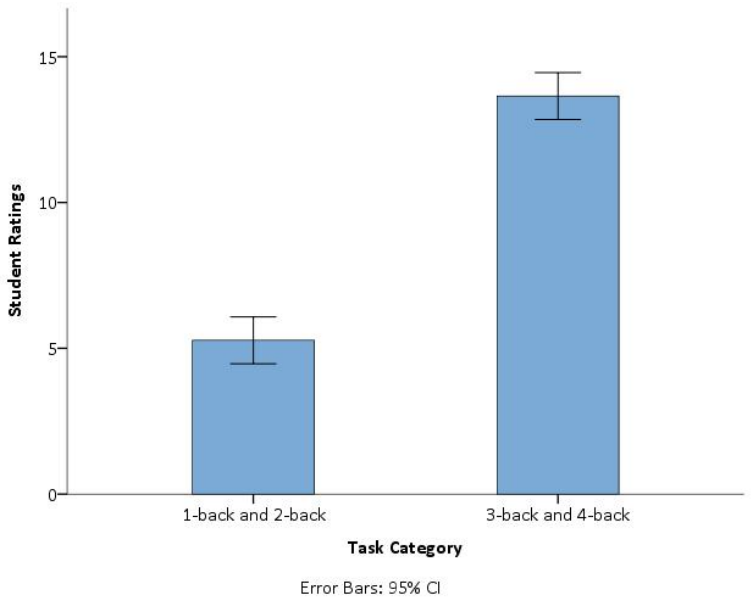

Figure 6: Combined mean student ratings for Question 1.

icant difference between the means of the two groups. The combined rating $R_{3,4}$ had a higher mean (M = 13.40, SD = 4.02) than $R_{1,2}$ (M = 5.52, SD = 3.14). However, the mean ratings were not enough to conclude that this difference is a real difference. Therefore, a paired-samples t-test was done to find the significance of the mean difference after removing one outlier from the sample(N =20). The test was found to be statistically significant, t(19) = -10.91, p < .001, d = 2.51. The results indicated that on average, students experienced more mental demand on the 3-back and 4-back tasks(M = 13.65, SD = 3.96) than on the 1-back and 2-back tasks(M = 5.28, SD = 3.00). The increment in mental demand for the two groups can be observed in Figure 6.

4.1.2 Question 2: How hurried or rushed was the pace of the task?

The second question asked students about their feeling regarding the speed/pace of the tasks. The mean ratings were, 1-back (M = 4.60, SD = 3.44), 2-back(M = 5.85, SD = 3.92), 3-back(M = 8.10, SD = 4.90), and 4-back(M = 8.40, SD = 3.93). Mean values of 3-back and 4-back tasks were almost similar. Overall, students did not feel that the tasks were too speedy. However, there was an increase in the mean ratings when the level of the n-back task increased. ANOVA test results indicated that there was a significant different between the four levels in their ratings, F(2.17, 41.16) = 16.26, p < .001, ω^2 = 1.59. However, there was no significant difference

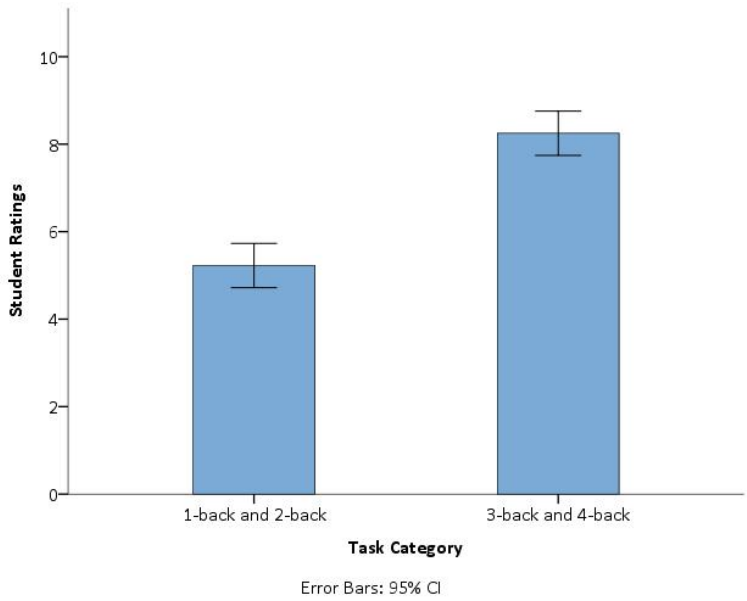

Figure 7: Combined mean student ratings for Question 2.

in the pace of tasks between 3-back and 4-back tasks, F(1,19) = .138, p = .715.

The mean combined ratings were calculated, and on average $R_{3,4}$ had a higher mean(M = 8.52, SD = 4.14) than $R_{1,2}$ (M =5.12, SD = 3.56). This increment of the mean ratings was not due to chance. The paired-samples t-test confirmed this, and there was a significant difference between $R_{1,2}$ (M=5.22, SD= 3.62) and $R_{3,4}$ (M = 8.25, SD = 4.05), t(19) = -6.24, p < .001, d = .82. The results indicated that students had to hurry up to finish the task, and it was observed highly in 3-back and 4-back tasks. In other words, their mental workload was higher during the last two levels of n-back tasks compared to the first two (see `Figure 7`).

4.1.3 Question 4: How hard did you have to work to accomplish your level of performance?

As I have not taken the 3^{rd} question into this analysis, let us move on to the 4^{th} question which asks about the effort students had to put in to find the best results. The mean ratings were as follows: 1-back (M = 4.95, SD = 4.36), 2-back (M = 7.62, SD = 3.50), 3-back(M = 12.52, SD = 4.28), and 4-back(M = 15.62, SD = 4.46). It was quite apparent that the mean ratings increased for higher levels of n-back tasks. ANOVA test proved this by showing significant difference between each level and the mean ratings for Question 4, F (1.94, 38.77) = 52.52, p < .001, $\omega^2 = 2.13$.

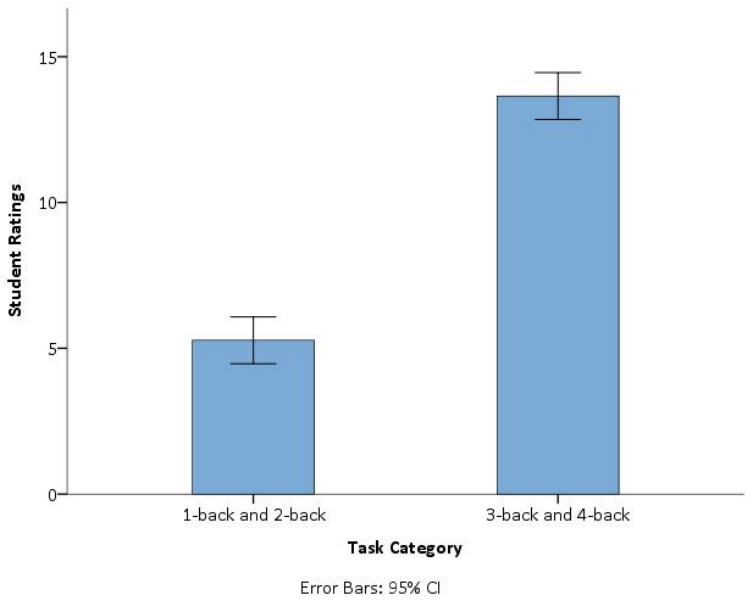

Figure 8: Combined mean student ratings for Question 4.

The paired-samples t-test conducted for $R_{1,2}$ (M = 6.28, SD = 3.80) and $R_{3,4}$ (M = 14.07, SD = 3.88) was also found statistically significant, t(20) = -8.66, p < .001, d = 2.05. Therefore, the mean difference between the two categories was a real difference, and the effort students put on to finish the task was higher when they reached higher levels of n-back tasks. During 1-back and 2-back tasks, students' effort was low compared to 3-back and 4-back tasks, thus, experiencing higher mental workload as shown in `Figure 8`.

4.1.4 Question 5: How insecure, discouraged, irritated, stressed, and annoyed were you?

This question can be considered as a direct indicator of mental workload. One outlier was removed from the sample to assume normality (N=20). Students reported a higher mean rating for 4-back task (M = 10.05, SD = 5.34) than 3-back(M = 7.35, SD = 4.52), 2-back(M = 4.80, SD = 4.03) and 1-back tasks(M = 2.90, SD = 2.67). In general, these values indicated that the metal workload gradually increased from 1-back to 4-back. This gradual increase was a real increase, and ANOVA test proved that by showing significant differences between the four levels in their ratings for Q5, F(1.92, 36.38) = 31.86, p < .001, ω^2 = 1.75.

The grouped mean ratings were calculated and the results showed that the mean difference ratings between $R_{1,2}$ (M = 3.85, SD = 3.23) and $R_{3,4}$ (M = 8.70,

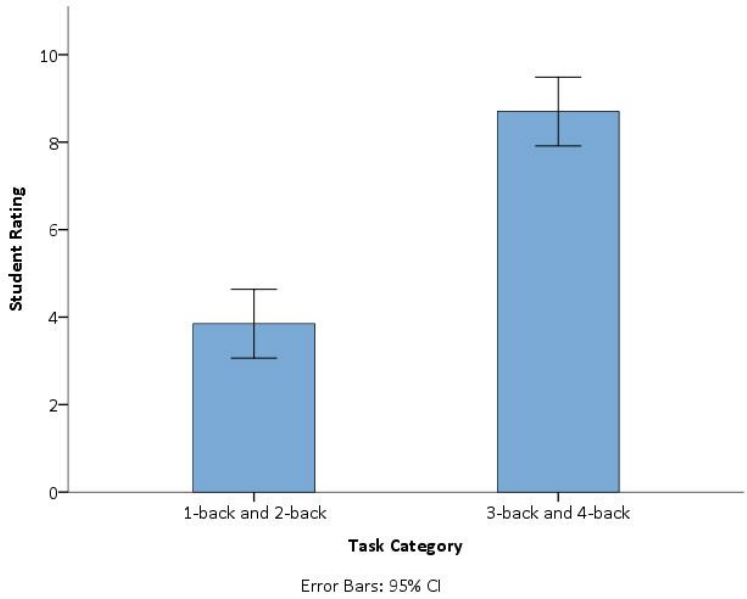

Figure 9: Combined mean student ratings for Question 5.

SD = 4.72) were statistically significant, t(19) = -6.46, p < .001 , d = 1.5. This is illustrated in Figure 9.

4.1.5 Categories of Mental Workload

Results of the ANOVA tests showed that there was a significant increase in the ratings with regards to the task level in all four questions. In addition to that, the comparison between *1-back and 2-back*, *2-back and 3-back*, and *3-back and 4-back* were found to be statistically significant. Therefore, we can conclude that the mental workload increased from 1-back to 4-back gradually. This is illustrated in the diagram shown in Figure 10.

Outcomes of the paired-samples t-tests described in the above sections lead us to divide the data into two categories of mental workload. $R_{1,2}$ in all the four questions was lower than $R_{3,4}$, and it was found that the difference between the two types was statistically significant. Therefore, we can conclude that the combined mental workload of 1-back and 2-back tasks was lower than the combined mean load of 3-back and 4-back tasks. For the simplicity of further analysis, we use these findings and group *1-back and 2-back* results into one category: *low mental workload* and the results of *3-back and 4-back* into another category: *high mental workload*. Figure 11 shows the increase of the mental workload for the combined groups.

Figure 10: Pattern of mental workload for the four n-back tasks.

Figure 11: Pattern of mental workload for the combined n-back tasks.

4.2 Eye Parameter Analysis

Both ANOVA test (together with the multivariate test) and paired-samples t-test were carried out for each of the 17 eye parameters, and the results for each eye parameter are presented in the following sections.

4.2.1 Blink Count

The sample (N=16) for the blink count analysis was taken by removing five participants who were considered as outliers. First, ANOVA test was carried out considering the four levels of n-back tasks. On average, blink count increased from 1-back to 3-back and decreased in the 4-back task. The values were as follows: 1-back (M = 40.29, SD = 23.61), 2-back(M = 55.29, SD = 27.48), 3-back(M = 78.00, SD = 57.45), and 4-back(M = 66.06, SD = 32.78). Overall, the difference between the blink count values for different mental workload levels was significant, F(1.24, 19.83) = 5.77, p = .021, ω^2 = .38. However, the individual difference in blink count between 2-back and 3-back was not significant, F(1, 16) = 3.18, p = .093. Similarly, no significant difference was found between 3-back and 4-back tasks, F(1,16) = .99, p = .334. However, in this scenario, multivariate results were considered as more powerful than the univariate results because Greenhouse-Geisser sphericity value was closer to its lower bound of .333(ε = .413)[30]. The multivariate test results indicated a significance in the differences, V = 0.64, F(3,14) = 8.40, p = .002, ω^2 = .38. Therefore, we can neglect the non-significant values for individual comparisons.

Next, the mean difference of blink count between low mental workload (M = 48.81, SD = 25.18) and high mental workload (M = 66.56, SD = 33.75) was

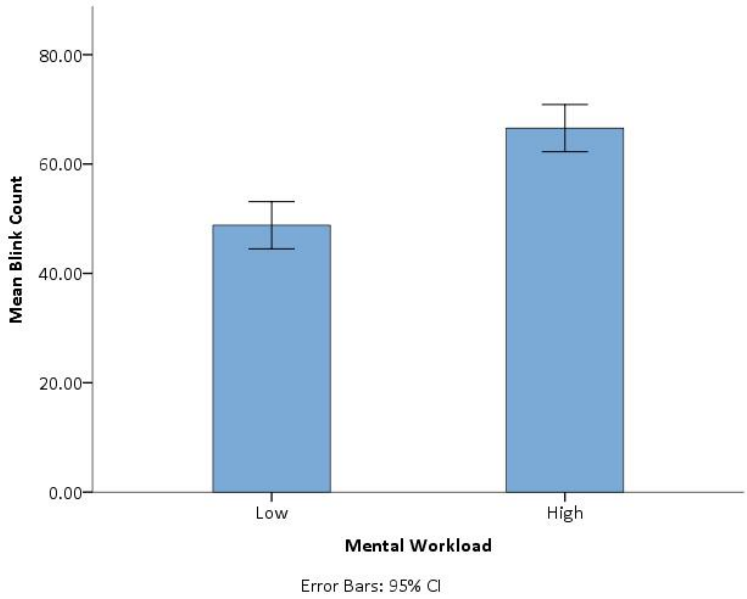

Figure 12: Mean blink count for low and high mental workload categories.

calculated. There was a strong positive correlation between the two categories, r(14) = .889, p < .001. From the paired-samples t-test, we can conclude that increased mental workload significantly affected their blinking (blink count), t(15) = -4.39, p = .001, d = .70. In other words, blink count tends to increase when students got more mental workload. This could be visualized in the bar graph shown in `Figure 12`. Finally, we can reject the null hypothesis (H_0), and accept the alternative hypothesis(H_1): *there is a significant difference in the "blink count" between students who experience "low mental workload" and "high mental workload"*.

4.2.2 Blink Frequency

Blink frequency is related to blink count, and it has a similar kind of behavior. Blink frequency is calculated as the blink count per millisecond. The same participants who were outliers became the outliers in blink frequency scores.

According to the descriptive statistics, following blink frequency values were found for the 4 tasks: 1-back (M = .32, SD = .20), 2-back (M = .45, SD = .23), 3-back(M = .62, SD = .46), and 4-back(M = .52, SD = .26). In the ANOVA test results, there was a significant difference between the four levels in their blink frequency values, F(1.21, 19.43) = 5.65, p = .023, ω^2= .36. The most suitable MANOVA test results showed that there was a significant difference in blink frequency between the four levels, V = 0.67, F(3, 14) = 9.58, p = .001, ω^2 = .36.

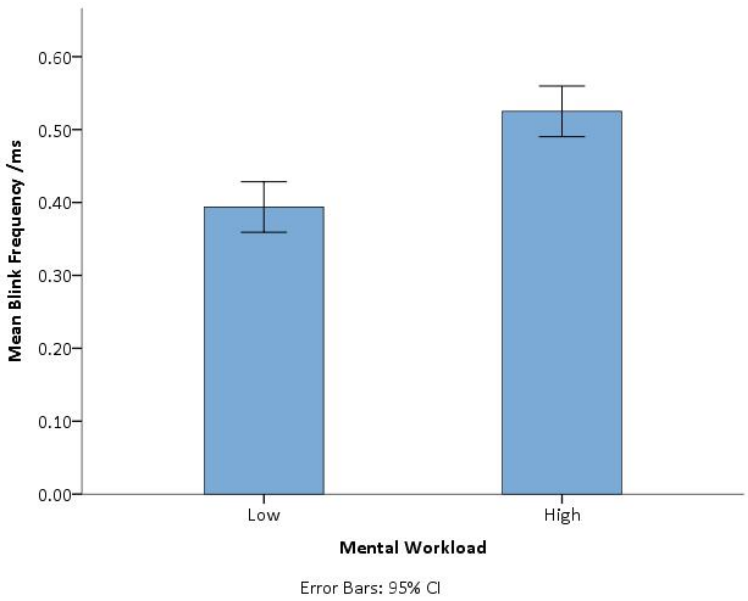

Figure 13: Mean blink frequency for low and high mental workload categories.

Similar to blink count, the individual difference in blink frequency between 2-back and 3-back was not significant, F(1, 16) = 2.87, p = .109. Same for 3-back and 4-back tasks, F(1,16) = 1.16, p = .297.

On average, students showed significantly higher blink frequency when they were subjected to high mental workload(M = .52 , SD = .28) than when they were subjected to low mental workload(M = .39, SD = .21), t(15) = -4.03 , p = .001, d = .62. The bar graph shown in `Figure 13` illustrates the respective mean blink frequencies between the two categories.

Considering both the test results, finally, we can reject the null hypothesis (H_0), and accept the alternative hypothesis(H_1): *There is a significant difference in the "blink frequency" between students who experience "low mental workload" and "high mental workload"*.

4.2.3 Blink Duration(Average)

Repeated-measures ANOVA test statistics were calculated to analyze the results of average blink duration. One outlier was removed to maintain normality(N = 20). The blink duration mean values and standard deviations for each task level were as follows: 1-back(M = 202.13, SD = 45.66), 2-back(M = 200.09, SD = 41.91), 3-back(M = 212.00, SD = 48.85), and 4-back(M = 227.50, SD = 60.48). A considerable increase in the mean values could be observed after the 2-back task.

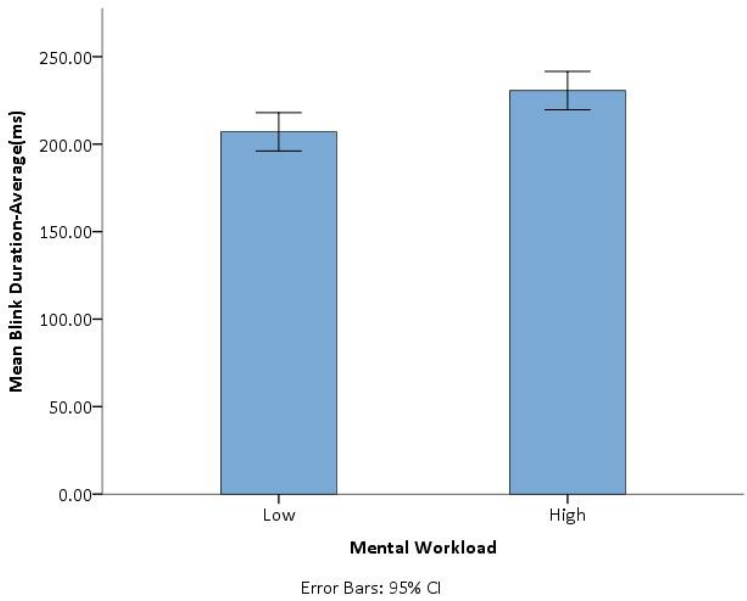

Figure 14: Mean blink duration(average) for low and high mental workload categories.

However, this increase was due to chance, and the difference was not significant according to the results of the ANOVA test, F(2.03, 38.57) = 1.88, p = .165, ω^2 = .14. MANOVA test results also indicated that the results were not significant, V = .22, F(3,17) = 1.63, p = .220, ω^2 = .14.

For average blink duration in the paired-samples t-test, no outliers were found. Therefore, the sample (N = 21) remained the original sample. Paired-samples t-test showed that the blink duration of a student was higher when imposed to high mental workload(M = 230.64, SD = 64.27), than when imposed to low mental workload(M = 207.08 , SD = 48.42), and the difference was significant, t(20)= -2.24, p= .037, d= .49. In addition to that, there was a positive correlation in blink duration between the two groups, r(14) = .668, p = .001. However, keep in mind that this difference was not very big if you observe the Cohen's effect value(d). `Figure 14` illustrates how low and high mental workloads affected the blink duration.

4.2.4 Peak Blink Duration

For the ANOVA test, 6 outliers had to be removed from the dataset (N = 15). Mean and standard deviations values were: 1-back (M = 453.72, SD = 401.03), 2-back (M = 573.89, SD = 248.81), 3-back (M = 574.97, SD = 283.94), and 4-back (M = 999.58, SD = 700.52). Mean value of 1-back has doubled when it comes to

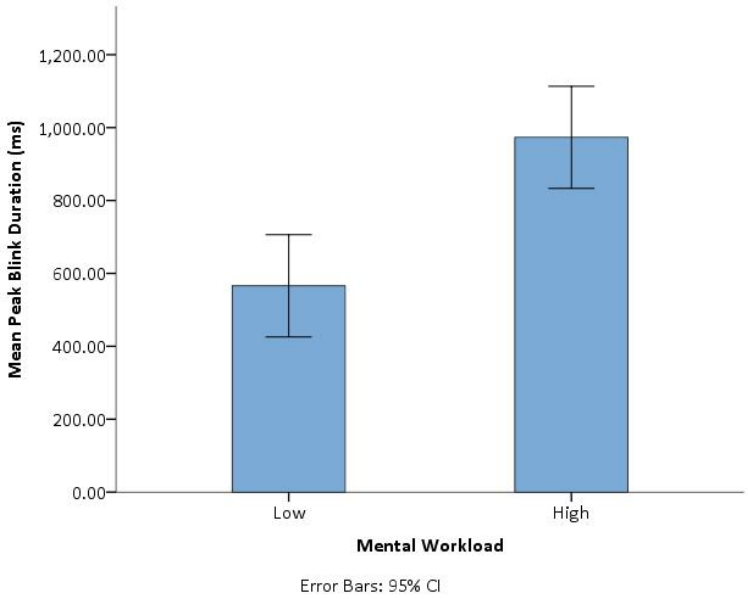

Figure 15: Mean peak blink duration(average) for low and high mental workload categories.

4-back. Results showed that there was a significant difference between the peak blink duration values in 4 n-back levels, $F(3,42) = 4.96$, $p < .001$, $\omega^2 = -4.52$. However, when considering the individual comparisons between each n-back level, only 3-back vs. 4-back levels had a significant difference, $F(1, 14) = 4.9$, $p = .044$. However, MANOVA test results indicated that there was no statistically significant difference in the mean peak blink duration values for the four n-back tasks, $V = .39$, $F(3,12) = 2.55$, $p = .105$, $\omega^2 = -4.52$. But in this scenario, ANOVA was considered as the most suitable reading as the sphericity was assumed according to Mauchly's test of sphericity($p = .061$) and Greenhouse-Geisser value(ε) was .70. Therefore, it was possible to neglect the MANOVA results.

For the paired-samples t-test, a normal distribution for mean differences of peak blink duration was obtained by removing four outliers ($N = 17$). The results showed that there was a significant difference in mean peak blink duration, $t(16) = -3.08$, $p = .007$, $d = 1.37$ between low workload condition($M = 565.94$, $SD = 296.74$) and high workload condition($M = 973.21$, $SD = 637.66$). In other words, when mental workload increased, peak blink duration would also increase, and it decreased under the low mental workload. This is shown in `Figure 15`.

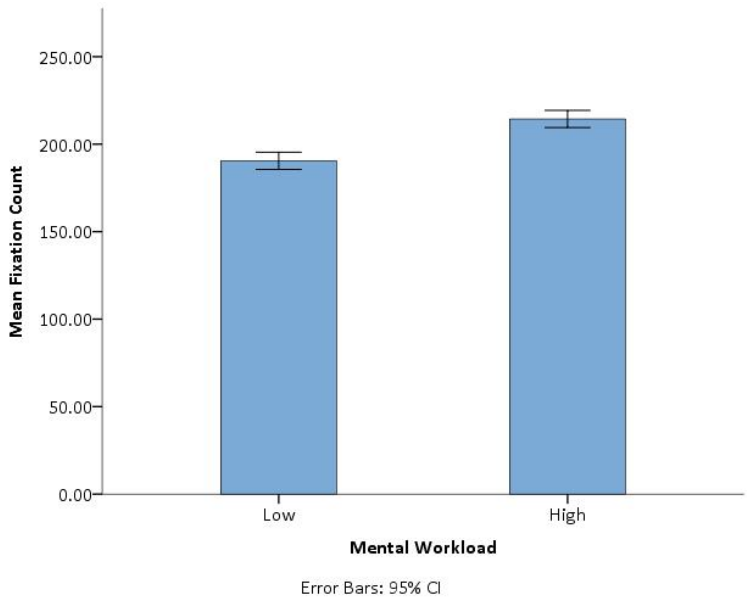

Figure 16: Mean fixation count for low and high mental workload categories.

4.2.5 Fixation Count

The ANOVA test was done using the complete dataset(N = 21). The results showed that the difference in mean fixation count between 1-back(M = 202.52, SD = 62.12), 2-back(M = 211.90, SD = 61.23), 3-back(M = 215.81, SD = 62.01), and 4-back(M = 207.48, SD = 61.78) was not significant, F(2.04, 40.90) = .24, p = .791, ω^2 = -.12. Furthermore, MANOVA test statistics also found to be not significant, V = 0.04, F(3, 18) = .26, p = .854, ω^2 = -.12. Therefore, it was clear that there was no significant difference in the fixation count between the four levels.

Next, fixation count analysis was done with a sample of 15 records after removing outliers for the low mental workload group and high mental workload group. The pattern for fixation count was also a positive pattern when it comes to higher mental workload as shown in `Figure 16`. The results indicated that the difference between low mental workload (M = 190.47, SD = 41.03) and high mental workload (M = 214.4 , SD = 38.83) was statistically significant, t(14) = -5.25, p < .001, d = .58 with regards to fixation count.

4.2.6 Fixation Frequency

Similarly, same sized samples(N = 21 and N = 15 respectively) were used to analyse fixation frequency. Descriptive statistics for the four n-back levels were as follows: 1-back(M = 1.62, SD = .50), 2-back(M = 1.70, SD = .49), 3-back(M =

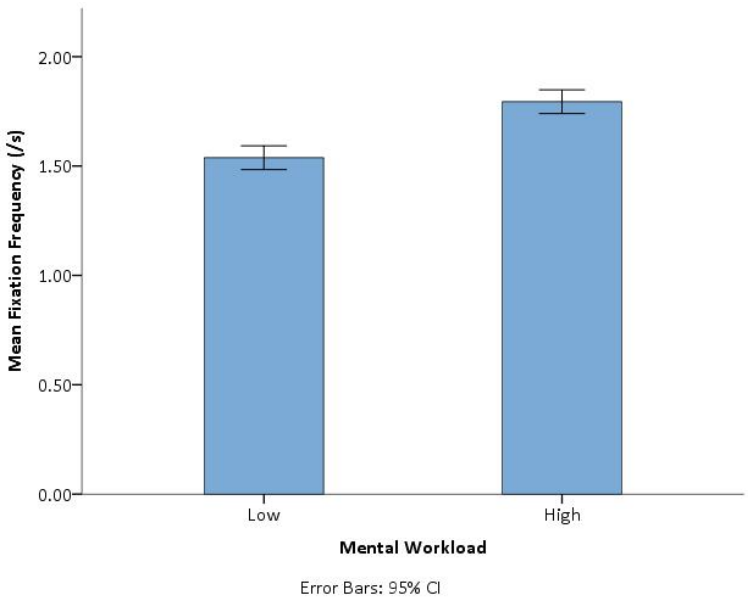

Figure 17: Mean fixation frequency for low and high mental workload categories.

1.73, SD = .49), and 4-back(M = 1.66, SD = .50). It was apparent that fixation frequency gradually increased up to 3-back task and then dropped in the 4-back task. ANOVA results showed that there was no significance in the difference between the fixation frequency values for different n-back tasks(i.e different mental workload levels), F(2.00, 40.08) = .28, p = .758, ω^2 = -.11. MANOVA results also proved that there was no significance between the two variables ,V = .05, F(3, 18) = .30, p = .822, ω^2 = -.11.

Similar to fixation count, fixation frequency showed a significant difference in mean for the t-test, t(16) = -4.99, p < .001, d = .81 between low mental workload(M = 1.54, SD = .31) and high mental workload (M = 1.79, SD = .35) as shown in Figure 17.

4.2.7 Fixation Duration

For the ANOVA test, no outliers were found (N = 21). The descriptive statistics were as follows: 1-back (M = 554.69, SD = 253.73), 2-back (M = 488.44, SD = 224.68), 3-back (M = 426.98, SD = 193.32), and 4-back (M = 498.61, SD = 429.12). These values did not have a clear pattern in the change in fixation duration, but there was a tendency to decrease. Results showed that fixation duration was not significantly affected by the different n-back levels (or mental workload levels), F(1.50,30.06) = .81, p = .423, ω^2 = -.02. The same was shown for the

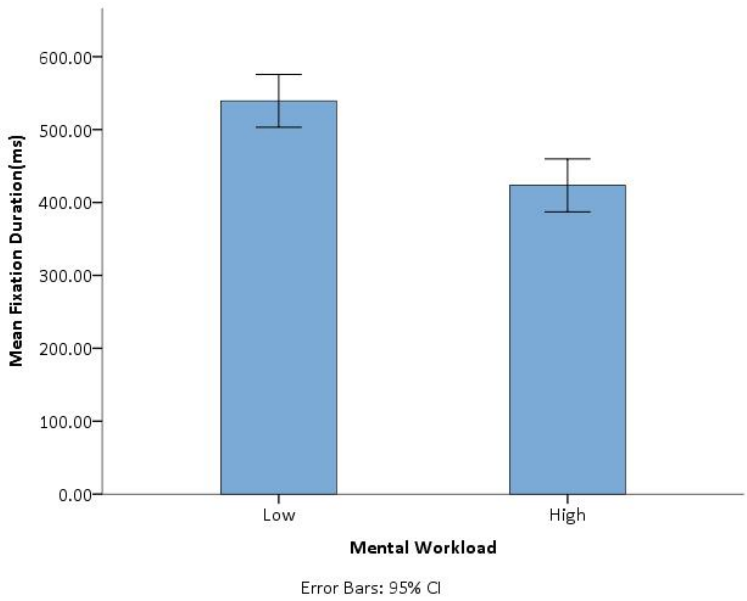

Figure 18: Mean fixation duration for low and high mental workload categories.

MANOVA test, and the results were not significant, V = .207, F(3,18) = 1.504, p = .233, ω^2 = -.02.

Removing one outlier, I got a sample of 20 students for the paired-samples t-test. Fixation duration was low when mental workload was high (M = 423.43, SD = 169.39) and high when you had a low mental workload (M = 539.37, SD = 207.5), and the difference was statistically significant, t(19) = 3.347, p = .003, d = .68. The more you fixate at a point, the less the mental workload and stress. `Figure 18` shows this relationship.

4.2.8 Peak Fixation Duration

Two outliers were removed from the original sample to achieve normality in the distribution (N = 19). ANOVA test showed that there was a significant difference in peak fixation duration between 1-back(M = 2950.24, SD = 502.85), 2-back(M = 2432.93 , SD = 540.32), 3-back(M = 2003.47, SD = 592.81) and 4-back(M = 2485.26, SD = 661.15) tasks, F(2.09, 37.69) = 10.52, p < .001, ω^2 = 1.20. The F-value was considerably high in *1-back vs. 2-back*(F(1,18) = 20.75, p < .001) than *2-back vs. 3-back*(F(1,18) = 12.86, p < .001) and *3-back vs. 4-back*(F(1, 18) = 6.12, p = .024). It was noted that the 4-back task had a sudden increase in peak fixation duration. MANOVA test was also found to be significant, V = 0.66, F(3,16) = 10.35, p < .001, ω^2 = 1.20.

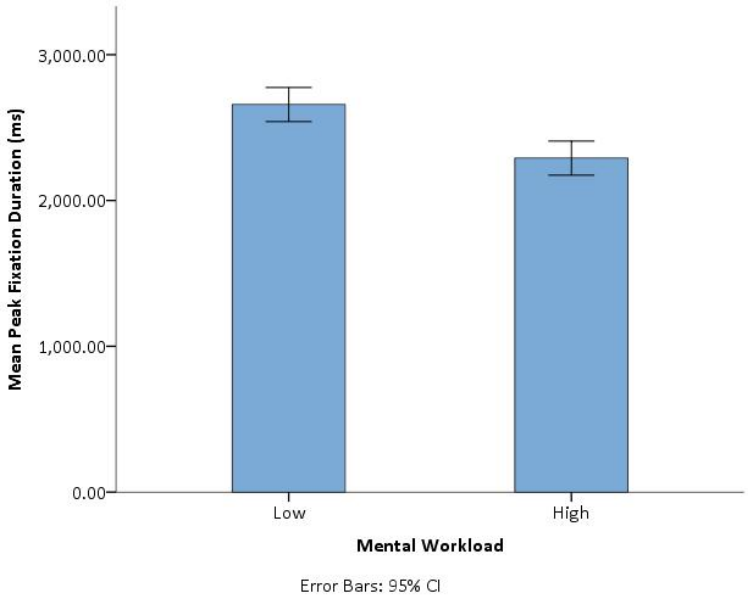

Figure 19: Mean peak fixation duration for low and high mental workload categories.

Peak fixation duration showed a negative trend compared to the other eye parameters as shown in Figure 19. Moreover, the mean difference between low mental workload (M = 2658.23, SD = 448.56) and high mental workload (M = 2290.81, SD = 428.02) was statistically significant, t(17) = 3.32 , p = .004, d = .86. This means that when users were highly mentally loaded they tend not to fixate a point for a long time. Therefore, the number of eye movements were higher than fixations.

Finally, it is possible to accept the alternative hypothesis, and conclude that there is a significant difference in the "peak fixation duration" between students who experience "low mental workload" and "high mental workload".

4.2.9 Saccade Count

More than 50% of the sample data had to be removed to assume the normality distribution of the mean difference of saccade count. Hence, the final sample consisted of just ten students. Due to this irregularity in the sample data(see Figure 20, this sample was not a good representation of the actual population. Regardless of the normality, the paired-samples t-test was performed for the original sample(N = 21). The sample mean and standard deviation values were as follows: 1-back(M = 251.29, SD = 218.28), 2-back(M = 217.62, SD = 151.61), 3-back(M = 252.76, SD = 206.56), and 4-back(M = 217.38, SD = 114.73). Results indi-

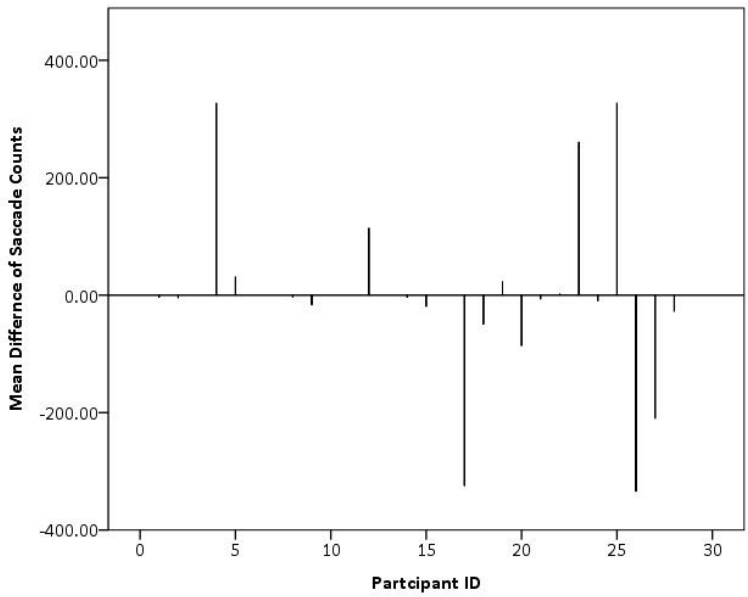

Figure 20: Distribution of mean saccade count difference for the sample(N= 21). Saccade counts were quite differenced from each participant.

cated that there was no significant difference on mean saccade count between the low mental workload(M = 234.45, SD = 168.59) and high mental workload(M = 235.07, SD = 146.84) , t(20) = -.02, p = .987, d = .00.

4.2.10 Saccade Frequency

Saccade frequency is the number of saccade counts per second. So, the same type of results like saccade count could be observed here also. The mean saccade frequency difference between low workload and high workload for all the participants are shown in `Figure 21`. The t-test results showed that the mean difference on saccade frequency between low mental workload(M = 1.87, SD = 1.35) and high mental workload(M = 1.88, SD = 1.17) was not statistically significant, t(20) = -.06, p = .956, d = .13.

4.2.11 Saccade Duration(Average)

Mean difference of saccade duration values between the two groups were found to be normally distributed for the original sample. Therefore, the paired-samples t-test was performed on this sample(N = 21). However, the results indicated that there was no statistically significant difference, t(20) =-.83, p = .415, d = .16, between the low mental workload(M = 45.54, SD = 5.84) and high mental workload(M =

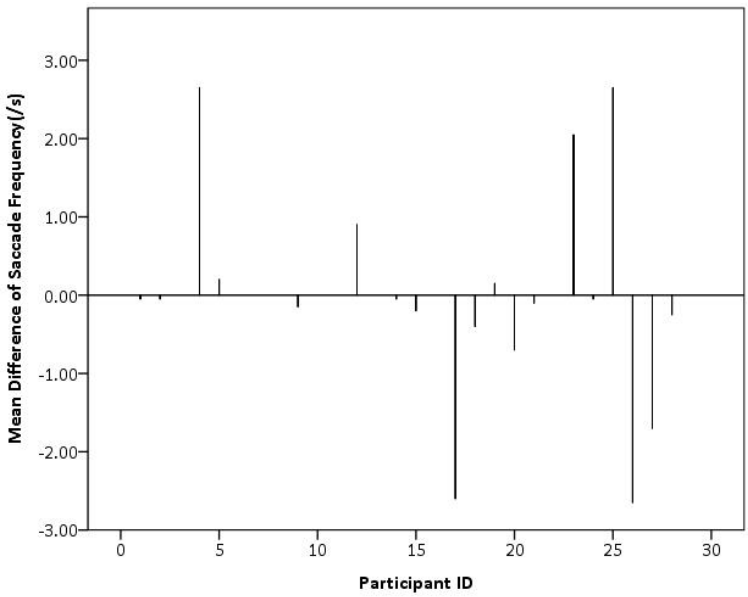

Figure 21: Distribution of mean saccade frequency difference for the sample(N= 21). Saccade frequencies were quite differenced from each participant.

46.50, SD = 4.56).

4.2.12 Peak Saccade Duration

Although mean peak saccade duration increased from low mental workload to high mental workload, it was due to chance. The t-test statistics proved this. The average difference of peak saccade duration between low mental workload(M = 116.19, SD = 39.74) and high mental workload(M = 124.82, SD = 63.00) were not statistically significant, t(19) = -.94, p = .360, d = .22. Therefore, the null hypothesis remains.

4.2.13 Saccade Amplitude

After removing six outliers, I got a normally distributed sample(N = 15) for saccade amplitude difference between low and high mental workload. It was observed a slight increase in saccade amplitude in high mental workload category(M = 4.40, SD = .68) than in low mental workload category(M = 4.34, SD = .57). However, this difference in mean saccade amplitude was not statistically significant, t(14) = -.93, p = .370, d = .09 according to the results found in paired-samples t-test.

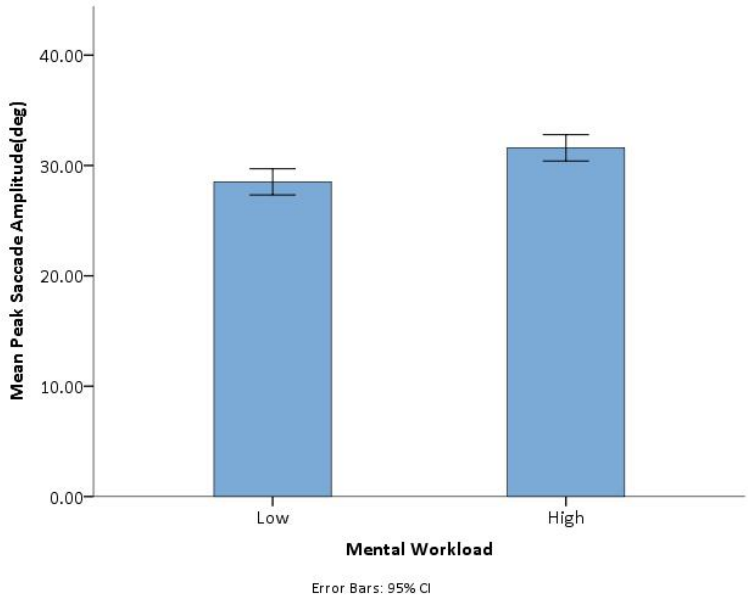

Figure 22: Mean peak saccade amplitude for low and high mental workload categories.

4.2.14 Peak Saccade Amplitude

Four outliers were removed from the original sample to satisfy the assumption of normality. The ANOVA test results indicated that there was no significant in the peak saccade amplitude between 1-back(M = 23.01, SD = 7.69), 2-back(M = 22.00, SD = 6.40), 3-back(M = 25.41, SD = 10.68), and 4-back(M = 27.86, SD = 9.76) tasks, F(3, 48) = 1.78, p = .164, ω^2 = -1.13. Moreover, there was no significant between the levels for the MANOVA test, V = 0.39 , F(3, 14) = 2.96 , p = .068 , ω^2 = -1.13.

Only one outlier was found in the original sample that prevented the normal distribution of the mean difference of peak saccade amplitude for the low and high mental workload groups. The sample(N = 20) showed a significant difference between low mental workload condition (M = 28.52, SD = 13.22) and high mental workload condition (M = 31.6, SD = 13.9), t(19) = -2.71, p = .014, d = .23. The trend is shown in `Figure 22`.

4.2.15 Saccade Velocity

The dataset for the ANOVA test was set by removing two outliers from the original dataset. The mean and standard deviation of the saccade velocity for the 4 levels were as follows: 1-back(M = 87.79, SD = 18.50), 2-back(M = 88.42, SD = 16.88), 3-back(M = 93.60, SD = 20.70), and 4-back(M = 101.20, SD = 29.67). However,

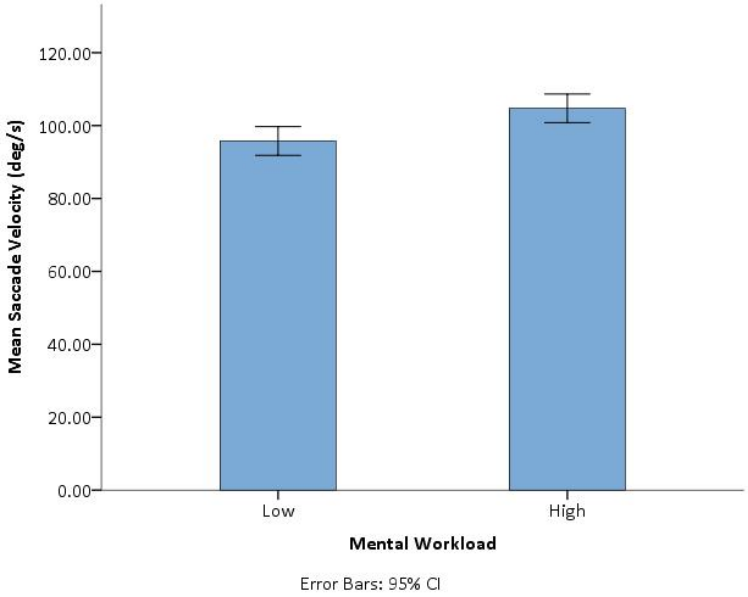

Figure 23: Mean saccade velocity for low and high mental workload categories.

this increase in the mean values was found to be non-significant, F(3, 54) = 1.91, p = .139, ω^2 = -1.39. Similarly, the MANOVA test results indicated that there was no significant relationship between the four mental workload levels, V = .25, F(3, 16) = 1.77, p = .194, ω^2 = -1.39.

There was just one outlier found related to the sample in saccade velocity(N=20). Similarly to most of the previous parameters, saccade velocity had a positive correlation to mental workload. The difference between mean saccade velocity for low workload(M = 95.78, SD= 36.91) and high workload(M = 104.74, SD = 37.46) was found to be significant, t(19) = -2.39, p = .027, d = .24. Therefore, when the participant's mental workload increased the saccade velocity also increased as shown in `Figure 23`.

4.2.16 Peak Saccade Velocity

However, peak saccade velocity cannot be used to measure mental workload like saccade velocity. It was observed during the t-test analysis that the mean difference between low mental workload(M = 585.40, SD = 150.80) and high mental workload(M = 588.99, SD = 140.49) was not statistically significant, t(20) = -.14, p = .893, d = .02.

4.2.17 Pupil Diameter

The original data sample was found to be normally distributed, and the ANOVA test was carried out on that sample. The results showed that the difference in mean pupil diameter between 1-back(M = 4.06, SD = .52), 2-back(M = 4.05, SD = .57), 3-back(M = 4.17, SD = .60), and 4-back(M = 4.14, SD = 14) was not significant, $F(1.54, 30.78) = 1.64$, $p = .212$, $\omega^2 = .06$. However, there was a significant difference in pupil diameter between 2-back and 3-back tasks, $F(1, 20) = 12.10$, $p = .002$. Due to the violation of sphericity($\varepsilon = .513$), test results of MANOVA can be considered over ANOVA. There was, in fact, a significant difference in the mean pupil diameter between the four levels according to the MANOVA test results, $V = 0.45$, $F(3,18) = 4.82$, $p = .012$, $\omega^2 = .06$.

Pupil diameter analysis using paired-samples t-test was done by filtering the sample removing two outliers and the new sample of 19 participants. There was a significant difference in means of pupil diameter, $t(18) = -4.568$, $p < .001$, $d = 0.24$ between low mental workload group (M = 4.09, SD = .55) and high mental workload group(M = 4.22, SD= .61). This means that when students had a higher mental workload, pupil diameter tends to increase than when they had a lower mental workload. Figure 24 shows this relationship, but the difference between the pupil diameter levels were relatively low (d = .24). However, the error bars are tiny, and they do not overlap, indicating that the significance between the two is still valid. Finally, we can reject the null hypothesis(H_0), and accept the alternative hypothesis(H_1): *there is a significant difference in the "pupil diameter" between students who experience "low mental workload" and "high mental workload"*.

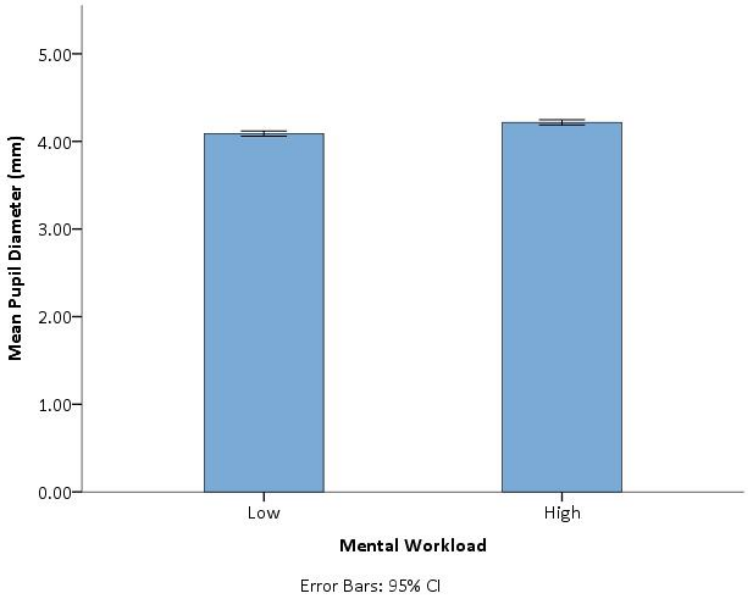

Figure 24: Mean pupil diameter for low and high mental workload categories.

5 Discussion

The two main purposes of this study were to find out which eye parameter has the most significant relationship with the mental workload of students and whether saccadic eye parameters have a significant influence on mental workload compared to blinks, fixations, and pupil diameter by using an n-back task which consisted of 4 levels. The study involved analyzing 17 eye parameters obtained from the *SMI red250mobile* eye tracker while performing an n-back task.

5.1 Analysis of the NASA-TLX Form Ratings

The f-value of the ANOVA test for Question 1 is a significantly larger value ($F(1.97, 37.35) = 88.22$). In addition to that, it has a large effect size of 2.53, and a significant probability of $p < .001$. Similarly, higher t-value($t(19) = -10.91$) and effect size($d = 2.51$) can be observed for the paired-samples t-test. It means that the mental demand experienced by students are vastly due to the manipulation of the four n-back tasks. Therefore, we can conclude that the experiment setup was successful, and it was able to generate higher mental demand on students at higher levels of the n-back tasks.

The ratings for the pace of the task are almost same for the 3-back task ($M = 8.10$, $SD = 4.90$), and 4-back($M = 8.40$, $SD = 3.93$) task. The results of the ANOVA level test (level 3 vs. level 4) statistically suggest this. Comparatively, both 3-back and 4-back tasks have a higher pace than 1-back and 2-back tasks. On the other hand, a significant difference can be seen in the speed of the task between the low and high mental workload categories in the paired-samples t-test. However, the pace of the task cannot be considered as a vital factor in the mental workload.

Students have to work harder in the higher levels of the n-back tasks compared to the lower levels of the n-back tasks. A higher f-value($F(1.94, 38.77) = 52.52$, $p < .001$) and effect size(2.13) in ANOVA test, and a higher t-value($t(20) = -8.66$) and effect size(2.05) in paired-samples t-test proves this. It is apparent that higher mental effort leads to higher mental workload[16].

Moreover, Question 4 which is a direct indicator of mental workload shows significant results for both ANOVA test and paired-samples t-test. Higher f-value($F(1.92, 36.38) = 31.86$), and t-value($t(19) = -6.46$), together with higher effect sizes: 1.75 and 1.5 respectively confirms this.

The analysis of the ratings of the four questions clearly indicates that there is a systematic increase in mental workload from 1-back task to the 4-back task, and

the users have experienced it throughout the experiment.

5.2 Analysis of the Eye Parameters

Analysis of the 17 eye parameters is discussed in the following sections. There are eye parameters which have shown statistically significant results both in ANOVA test and paired-samples t-test. Those parameters can be considered as the most suitable parameters to estimate mental workload. In addition to that, some eye parameters show a significant relationship to the mental workload only in paired-samples t-test, whereas some others do not show any significant relationship in both the tests. The former can be used to estimate mental workload to some extent(partially), but they are not very strong, while the latter cannot be used to estimate mental workload.

5.2.1 Blink Count and Blink Frequency

Based on the ANOVA and MANOVA results, blink count increases significantly when mental workload increases. Effect size for ANOVA is medium ($\omega^2 = .38$) and large for paired-samples t-test($d = .70$). Multivariate f-value is a relatively larger value ($F(3,14) = 8.40$) and it is significant at $p = .002$. Therefore, it is clear that there is a significant difference in blink count between students who experience low mental workload and high mental workload.

Almost similar results can be seen for blink frequency, where effect sizes for ANOVA test and t-test are .36 and .62 respectively. Both are causing a medium effect with regards to mental workload. However, the f-value for the MANOVA test ($F(3, 14) = 9.58$) is higher than that of the blink count. Therefore, blink frequency proves to be a better parameter than blink count in estimating mental workload. This is in line with the literature where blink frequency is favored over blink count. However, the results of the tests, where blink frequency increases with the mental workload prove Tsai et. al[6] 's findings but in contradict with Ledger[31] and Zheng et al.[29].

5.2.2 Blink Duration

Although paired-samples t-test showed a significant relationship between blink duration and mental workload with a medium effect, ANOVA test found to be non-significant. Although the t-test was significant, its significant probability($p = .037$) is closer to .05 margin. Furthermore, MANOVA test was not significant either. Therefore, it is not possible to conclude a significant relationship merely based on the t-value.

5.2.3 Peak Blink Duration

Both the ANOVA test and t-test shows a significant relationship with the mental workload. A relatively strong relationship can be observed between 3-back and 4-back tasks. However, there is a positive correlation where peak blink duration increases with the mental workload. This is a new finding, and no research has previously focused on the relationship between peak blink duration and mental workload.

5.2.4 Fixation Count and Fixation Frequency

However, both fixation count and fixation frequency do not have strong evidence to be called as useful parameters to estimate mental workload. They are suitable, when we use the combined mental workloads according to t-test results, but not under individual n-back tasks. This can be seen in ANOVA and MANOVA test results. Both show non-significant differences in fixation count and fixation frequency for all four levels of n-back tasks. The previous study of Wang et al.[32] also has failed to show a significant relationship of fixation count/fixation frequency to the mental workload. However, if we consider the first 3-tasks (1, 2 and 3-back tasks), there is an increase in fixation frequency with the increase in mental workload. However, at the 4-back level, sudden decrease in fixation frequency can be observed. It is possible to think that this situation as a mentally overloaded situation. Therefore, the results are exactly representing the same pattern of He et al.[23], where fixation frequency increases with mental workload and decreases when mentally overloaded. The same trend can be observed in fixation count. These are shown in `Figure 25` and `Figure 26` respectively.

5.2.5 Fixation Duration

Both the ANOVA and MANOVA tests showed non-significant results for fixation duration. However, the combined mental workload groups showed significant differences. Nevertheless, it is not enough to conclude the relationship between the two variables. However, if we assume that 4-back task is an overloaded point, then we can say that fixation duration decreases with mental workload and increases when the student is mentally overloaded.

5.2.6 Peak Fixation Duration

Both the ANOVA test, MANOVA test, and the t-test shows a significant relationship with the mental workload. The peak fixation duration tends to decrease when the mental workload increases. A relatively strong relationship can be observed between 1-back and 2-back tasks. Peak fixation duration has a more robust relationship with mental workload than peak blink duration. For example, peak blink duration has an f-value of $F(3, 42) = 4.96$ and effect size of -4.52, whereas peak

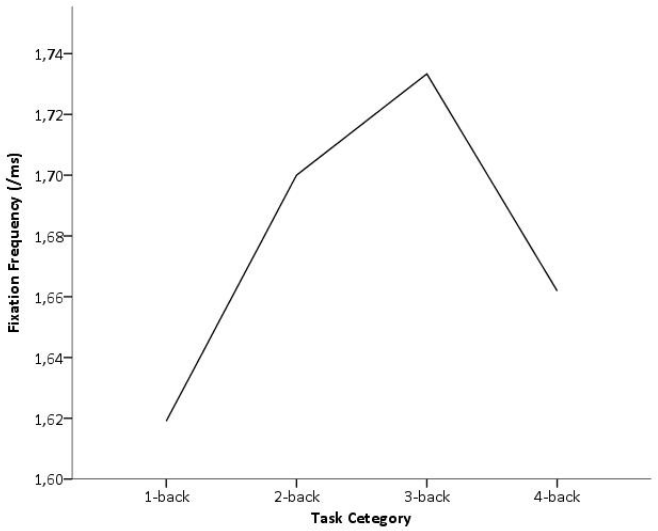

Figure 25: Mean fixation frequency of the four n-back tasks. Sudden fall in the frequency can be observed in the 4-back task.

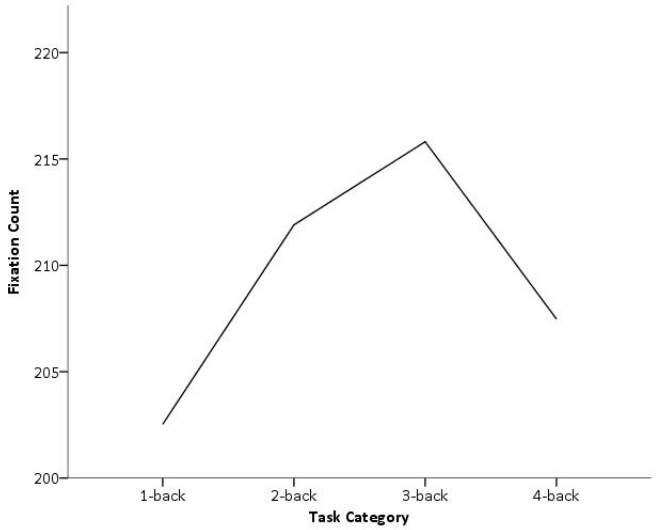

Figure 26: Mean fixation count of the four n-back tasks. Sudden fall in the values can be observed in the 4-back task.

fixation duration has a higher f-value F(2.09, 37.69) = 10.52 and effect size, 1.20. This is a new finding, and no research has previously focused on the relationship between peak fixation duration and mental workload.

5.2.7 Peak Saccade Amplitude

Both ANOVA and MANOVA tests fail to find a significant relationship between peak saccade amplitude and mental workload. However, the combined mental workload groups find a significant relationship between the two groups.

5.2.8 Saccade Velocity

Similarly, both ANOVA and MANOVA tests fail to find a significant relationship between saccade velocity and mental workload. However, the combined mental workload groups find a significant relationship between the two groups. Therefore, it can be used to estimate mental workload partially to some extent.

5.2.9 Pupil Diameter

MANOVA test results indicate that there is a significant relationship between pupil diameter and the mental workload. Moreover, significant t-test value supports this claim. However, the effects are quite small, and they are $\omega^2 = .06$, and d = 0.24 respectively. In general, pupil diameter increases with the mental workload[3]. The validity of this can be assured using the results of this study, where it shows a significant increase in pupil diameter when the mental workload increases, both in MANOVA test($V = 0.45$, $F(3,18) = 4.82$, $p = .012$, $\omega^2 = .06$.) and t-test($t(18) = -4.57$, $p < .001$, $d = 0.24$).

5.2.10 Non-significant Parameters

Saccade count sample consists of various values and dispersed. Paired-samples t-test values show that there is no significant relationship with the mental workload. Similarly, saccade frequency, average saccade duration, peak saccade duration, saccade amplitude, and peak saccade velocity have no relationship with the mental workload. All the non significant parameters are saccade parameters. However, in literature, there are studies that have shown significant relationships between workload and some of these parameters. External factors in addition to the mental workload might have affected the results of the study. However, this needs to be further investigated in future research.

5.2.11 Summary

The above parameters can be sorted based on their ability to estimate mental workload. This is illustrated in `Table 1`. According to the table, peak fixation duration can be considered as the most suitable eye parameter to estimate mental workload of university students.

Parameter	**t-test**	**ANOVA**	**MANOVA**	**Relationship**
Peak fixation duration	t(17)=3.32, **p=.004**, d=.86	F(2.09,37.69)=10.52, **p<.001**, ω^2=1.20	V=.66, F(3,16)=10.35, **p<.001**, ω^2=1.20	YES
Blink frequency	t(15)=-4.03, **p=.001**, d=.62	F(1.21,19.43)=5.65, **p=.023**, ω^2=.36	V=.67, F(3,14)=9.58, **p=.001**, ω^2=.36	YES
Blink count	t(15)=-4.39, **p=.001**, d=.70	F(1.24,19.83)=5.77, **p=.021**, ω^2=.38	V=.64, F(3,14)=8.40, **p=.002**, ω^2=.38	YES
Peak blink duration	t(16)=-3.08, **p=.007**, d=1.37	F(3,42)=4.96, **p<.001**, ω^2=-4.52	V=.39, F(3,12)=2.55, p=.105, ω^2=-4.52	YES
Pupil diameter	t(18)=-4.57, **p<.001**, d=.24	F(1.54,30.78)=1.64, p=.212, ω^2=.06	V=.45, F(3,18)=4.82, **p=.012**, ω^2=.06	YES
Fixation frequency	t(16)=-4.993, **p<.001**, d=.81	F(2.00,40.08)=.28, p=.758, ω^2=-.11	V=0.05, F(3,18)=.30, p=.822, ω^2=-.11	PARTIAL
Fixation duration	t(19)=3.347, **p=.003**, d=.68	F(1.50,30.06)=.81, p=.423, ω^2=-.02	V=0.21, F(3,18)=1.50, p=.233, ω^2=-.02	PARTIAL
Fixation count	t(14)=-5.247, **p<.001**, d=.58	F(2.04,40.90)=.24, p=.791, ω^2=-.12	V=.04, F(3,18)=.26, p=.854, ω^2=-.12	PARTIAL
blink duration	t(20)=-2.241, **p=.037**, d=.49	F(2.03,38.57)=1.88, p=.165, ω^2=.14	V=.22, F(3,17)=1.63, p=.220, ω^2=.14	PARTIAL
Saccade velocity	t(19)=-2.390, **p=.027**, d=.24	F(3,54)=1.91, p=.139, ω^2=-1.39	V=.25, F(3,16)=1.77, p=.194, ω^2=-1.39	PARTIAL
Peak saccade amplitude	t(19)=-2.711, **p=.014**, d=.23	F(3,48)=1.78, p=.164, ω^2=-1.13	V=.39, F(3,14)=2.96, p=.068, ω^2=-1.13	PARTIAL
Peak saccade duration	t(19)=-.938, p=.360, d=.22	-	-	NO
Saccade duration (Avg.)	t(20)=-.832, p=.415, d=.16	-	-	NO
Saccade frequency	t(20)=-.056, p=.956, d=.13	-	-	NO

Saccade amplitude	t(14)=-.926, p=.370, d=.09	-	-	NO
Peak saccade velocity	t(20)=-.137, p=.893, d=.02	-	-	NO
Saccade count	t(20)=-.017, p=.987, d=.00	-	-	NO

Table 1: Summary of the test results

6 Conclusions and Future Work

This chapter will present the conclusions drawn from the conducted research, together with the suggested future work.

6.1 Conclusions

Based on the analysis of the NASA-TLX form ratings and the eye parameters, following conclusions can be made.

- As many other studies have suggested, visual n-back tasks can be used to increase a person's mental workload systematically, and it was observed among the students during the study. Therefore, using the n-back task to induce the mental workload can be considered as a valid method.
- As stated in the beginning, one of the principal purposes of this research was to find out the most significant eye parameter that can be used to estimate mental workload of university students. Results suggest that *peak fixation duration* has a strong negative significant relationship with the mental workload and it is the most reliable eye parameter that can be used to estimate mental workload among the 17 eye parameters.
- Both *blink count* and *blink frequency* have a significant positive relationship with the mental workload, but *blink frequency* has the strongest relationship of the two. *Blink frequency* results confirm the validity of previous research findings. *Blink count* is associated with *blink frequency*. Therefore, it is harder to find any study that has discussed separately on *blink count*. That gap is filled by the findings in this study.
- *Peak blink duration* shows a positive relationship with the mental workload. In other words, *peak blink duration* increases with the mental workload. This is a new finding related to estimating mental workload.
- The results on *pupil diameter* which show a significant positive relationship with mental workload confirms the validity of earlier findings. Though lighting conditions can increase pupil diameter, in this experiment, it was not affected because of the same lighting conditions used for every user. Therefore, the increased value of pupil diameter due to lighting conditions can be kept as constant for every user, and it does not have any negative impact on our conclusions.
- The research approach of this analysis is unique. Some eye parameters showed

significant results in the t-test, but not in ANOVA and MANOVA test. Therefore, *fixation frequency*, *fixation duration*, *fixation count*, *blink duration*, *saccade velocity*, and *peak saccade amplitude* are concluded as partially supported parameters for estimating mental workload. According to t-test results, fixation duration has a negative relationship with mental workload whereas the other partially supported parameters have positive relationships.

- Among 17 eye parameters, *peak saccade duration*, *saccade duration(avg.)*, *saccade frequency*, *saccade amplitude*, *peak saccade velocity*, and *saccade count* have no significant relationship with the mental workload. According to the results of the thesis, these parameters cannot be used to estimate the mental workload of students.
- Measurements of blinks, fixations and pupil diameter are more reliable parameters for estimating mental workload than saccade eye parameters. These findings are useful to develop software that might detect the mental workload by measuring the eye parameters and notify the users once they get overloaded.

6.2 Future Work

The results of this research lead to some interesting future work. These are addressed in the following.

- Those collected data can be analyzed further to find any gender effect on the mental workload of the students. It is interesting to see which gender has the highest possibility to get mentally overloaded. In the same way, it is possible to see the effect of previous experience to mental workload estimation. For example, new users and previously participated users in the pilot test can be used to see whether the experience in the n-back task would have any effect on eye parameters with regards to mental workload.
- This study can be extended to find a way to investigate the behavior of saccade intrusion with the existing data. It is well known as a useful parameter to estimate mental workload since lighting conditions do not affect the results of saccadic intrusions unlike in pupil diameter[11].
- It is interesting to find by giving a user the simplest task first and then the most challenging task(i.e., 4-back task) and see how the eye parameters behave. It will help to identify how eye parameters behave when the unexpected mental workload is applied. In addition to that, it is motivating to conduct the same experiment for different user groups using larger samples.
- The images used in the study have different brightness and luminance levels based on the colors used. Therefore, it is interesting to conduct the same experiment using only black and white images and see whether same results

can be obtained with them.

Bibliography

[1] Jorna, P. G. 1992. Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological psychology*, 34(2), 237–257.

[2] Brookhuis, K. A., De Waard, D., Kraaij, J. H., & Bekiaris, E. 2003. How important is driver fatigue and what can we do about it. *Human Factors in the Age of Virtual Reality. Maastricht: Shaker Publishing*, 191, 207.

[3] Gable, T. M., Kun, A. L., Walker, B. N., & Winton, R. J. 2015. Comparing heart rate and pupil size as objective measures of workload in the driving context: initial look. In *Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 20–25. ACM.

[4] Chen, S. & Epps, J. 2014. Using task-induced pupil diameter and blink rate to infer cognitive load. *Human–Computer Interaction*, 29(4), 390–413.

[5] Heeman, P. A., Meshorer, T., Kun, A. L., Palinko, O., & Medenica, Z. 2013. Estimating cognitive load using pupil diameter during a spoken dialogue task. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 242–245. ACM.

[6] Tsai, Y.-F., Viirre, E., Strychacz, C., Chase, B., & Jung, T.-P. 2007. Task performance and eye activity: predicting behavior relating to cognitive workload. *Aviation, space, and environmental medicine*, 78(5), B176–B185.

[7] Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., & Montanari, R. 2011. Driver workload and eye blink duration. *Transportation research part F: traffic psychology and behaviour*, 14(3), 199–208.

[8] Bodala, I. P., Kukreja, S., Li, J., Thakor, N. V., & Al-Nashash, H. 2015. Eye tracking and eeg synchronization to analyze microsaccades during a workload task. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, 7994–7997. IEEE.

[9] Chandler, P. & Sweller, J. 1991. Cognitive load theory and the format of instruction. *Cognition and instruction*, 8(4), 293–332.

[10] Paas, F. G. & Van Merriënboer, J. J. 1994. Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of educational psychology*, 86(1), 122.

[11] Tokuda, S., Obinata, G., Palmer, E., & Chaparro, A. 2011. Estimation of mental workload using saccadic eye movements in a free-viewing task. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, 4523–4529. IEEE.

[12] Di Stasi, L. L., Antolí, A., & Cañas, J. J. 2011. Main sequence: an index for detecting mental workload variation in complex tasks. *Applied ergonomics*, 42(6), 807–813.

[13] Di Stasi, L. L., Álvarez-Valbuena, V., Cañas, J. J., Maldonado, A., Catena, A., Antolí, A., & Candido, A. 2009. Risk behaviour and mental workload: Multimodal assessment techniques applied to motorbike riding simulation. *Transportation research part F: traffic psychology and behaviour*, 12(5), 361–370.

[14] Marquart, G., Cabrall, C., & de Winter, J. 2015. Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*, 3, 2854–2861.

[15] Ohtsuka, R., Chihara, T., Yamanaka, K., Morishima, K., Daimoto, H., et al. 2015. Estimation of mental workload during motorcycle operation. *Procedia Manufacturing*, 3, 5313–5318.

[16] Paas, F. G. & Van Merriënboer, J. J. 1994. Instructional control of cognitive load in the training of complex cognitive tasks. *Educational psychology review*, 6(4), 351–371.

[17] Parasuraman, R. & Caggiano, D. 2002. Mental workload. *Encyclopedia of the human brain*, 3, 17–27.

[18] Di Stasi, L., Marchitto, M., Antolí, A., & Cañas, J. 2013. Saccadic peak velocity as an alternative index of operator attention: A short review. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, 63(6), 335–343.

[19] Bodala, I. P., Ke, Y., Mir, H., Thakor, N. V., & Al-Nashash, H. 2014. Cognitive workload estimation due to vague visual stimuli using saccadic eye movements. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, 2993–2996. IEEE.

[20] Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. 2011. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

[21] Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. 2010. The concurrent validity of the n-back task as a working memory measure. *Memory*, 18(4), 394–412.

[22] Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, 38(1), 63–71.

[23] He, X., Wang, L., Gao, X., & Chen, Y. 2012. The eye activity measurement of mental workload based on basic flight task. In *Industrial Informatics (INDIN), 2012 10th IEEE International Conference on*, 502–507. IEEE.

[24] Hart, S. G. & Staveland, L. E. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52, 139–183.

[25] 2010. *image of the paper-and-pencil version of the nasa-tlx rating scale*. NASA. URL: https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLXScale.pdf.

[26] Pfleging, B., Fekety, D. K., Schmidt, A., & Kun, A. L. 2016. A model relating pupil diameter to mental workload and lighting conditions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5776–5788. ACM.

[27] Bierbaum, C. R., Szabo, S. M., & Aldrich, T. B. 1989. *Task analysis of the UH-60 mission and decision rules for developing a UH-60 workload prediction model: Summary report*, volume 1. US Army Research Institute for the Behavioral and Social Sciences.

[28] Schneider, M. & Deml, B. 2016. An integrated approach of mental workload assessment. In *Advances in Ergonomic Design of Systems, Products and Processes*, 191–208. Springer.

[29] Zheng, B., Jiang, X., Tien, G., Meneghetti, A., Panton, O. N. M., & Atkins, M. S. 2012. Workload assessment of surgeons: correlation between nasa tlx and blinks. *Surgical endoscopy*, 26(10), 2746–2750.

[30] Field, A. 2013. *Discovering statistics using IBM SPSS statistics*. Sage.

[31] Ledger, H. 2013. The effect cognitive load has on eye blinking. *The Plymouth Student Scientist*, 6(1), 206–223.

[32] Wang, Q., Yang, S., Liu, M., Cao, Z., & Ma, Q. 2014. An eye-tracking study of website complexity from cognitive load perspective. *Decision support systems*, 62, 1–10.

[33] Schulz, C., Schneider, E., Fritz, L., Vockeroth, J., Hapfelmeier, A., Wasmaier, M., Kochs, E., & Schneider, G. 2011. Eye tracking for assessment of workload: a pilot study in an anaesthesia simulator environment. *British journal of anaesthesia*, 106(1), 44–50.

[34] Cardona, G. & Quevedo, N. 2014. Blinking and driving: the influence of saccades and cognitive workload. *Current eye research*, 39(3), 239–244.

[35] Di Stasi, L. L., Marchitto, M., Antolí, A., Baccino, T., & Cañas, J. J. 2010. Approximation of on-line mental workload index in atc simulated multitasks. *Journal of Air Transport Management*, 16(6), 330–333.

[36] Beatty, J. & Lucero-Wagoner, B. 2000. The pupillary system, handbook of psychophysiology, cacioppo, tassinary & berntson.

[37] Beatty, J. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin*, 91(2), 276.

[38] Cao, Y., Kobayashi, Y., Zhang, B., Liu, Q., Sugano, S., & Fujie, M. G. 2015. Evaluating proficiency on a laparoscopic suturing task through pupil size. In *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, 677–681. IEEE.

[39] Edirisinghe, V. R. D. A. 2017. N-back website. http://nback.azurewebsites.net. Accessed: 2017-03-30.

[40] Edirisinghe, V. R. D. A. 2017. N-back practice test. http://nback.azurewebsites.net/home/welcome. Accessed: 2017-03-30.

[41] SensoMotoricInstruments. 2017. Smi red250mobile. https://www.smivision.com/eye-tracking/product/red250mobile-eye-tracker. Accessed: 2017-04-10.

A Screen shots of the n-back Experiment Website

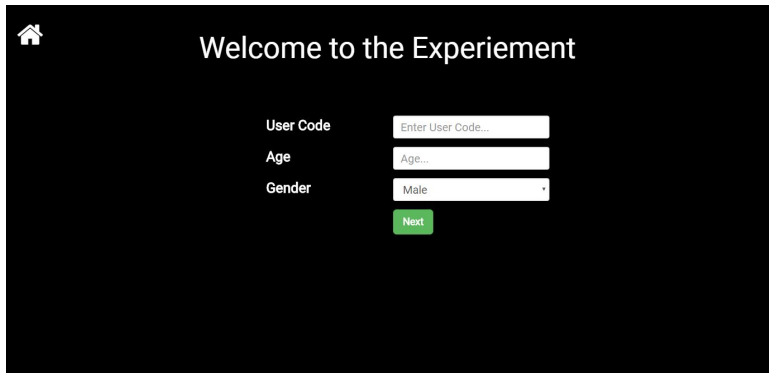

Figure A.0.1: Login screen.

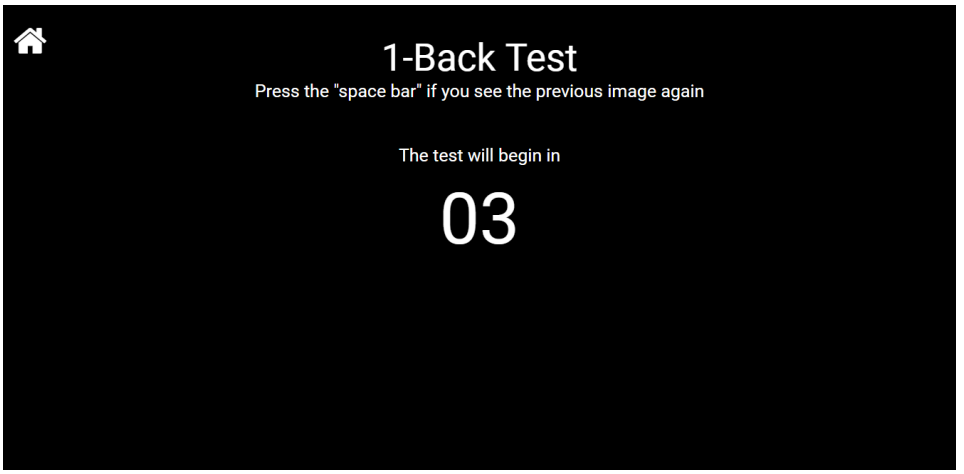

Figure A.0.2: Sample page 1

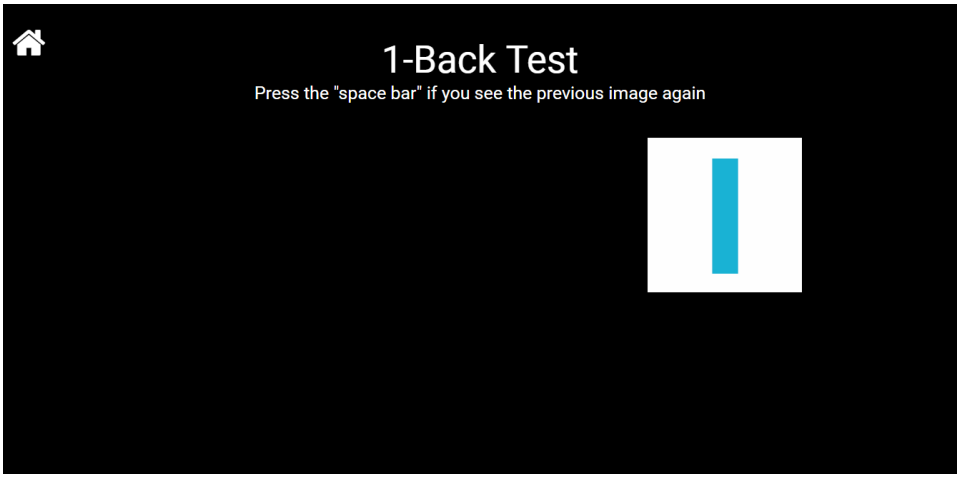

Figure A.0.3: Sample page 2

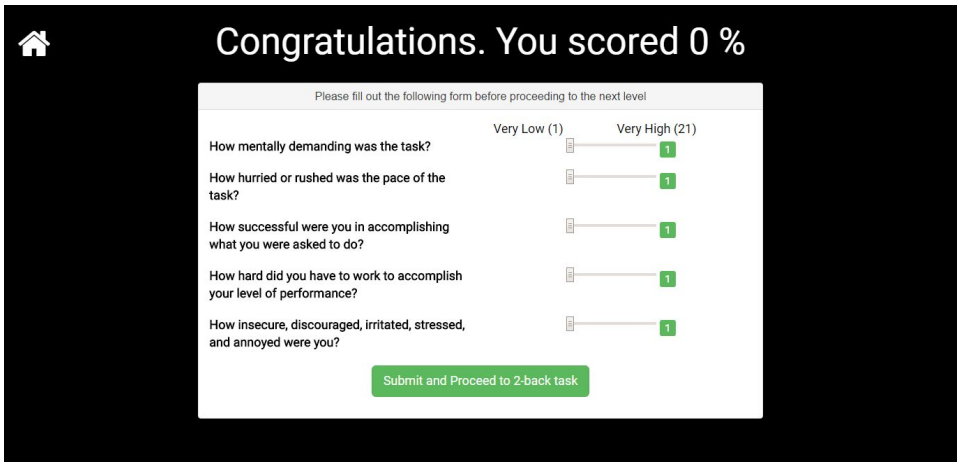

Figure A.0.4: Login screen 3.

Figure A.0.5: Sample page 4.

Figure A.0.6: Sample page 6.

B Informed Consent Form and Instructions

Consent form for participation in an eye-tracker experiment for the Master Thesis project

Background and Purpose
The goal of this research experiment is, to investigate the mental workload using different types of eye parameters. This includes doing a mixed-methods research using a remote eye tracker as a tool for measuring. This project is conducted for the Thesis, which is part of the Master in Interaction Design program at NTNU. The sample for this experiment is selected using convenience sampling technique. That is by asking the students or other workers who are around in the university premises, the possibility of taking part in the experiment.

What does participation in the project imply?
Participants will get a brief explanation of what they are expected to do before starting the test. Users will get a sequence of images, where they supposed to remember 1 image back, 2 images back, 3 images back and 4 images back. Users will get a practice session to get a good understanding of what they supposed to do before they conduct the real experiment. The whole experiment will take approximately 15-20 minutes. There will not be any audio recordings. Only eye movements will be recorded. No sensitive information will be gathered during the experiment, and any information you provide will have a number instead of your name. Every participant will get an evaluation form after completing each task before they go to the next level.

What will happen to the collected information?
The information gathered during this experiment will exclusively be used for the described school project, and that information is only accessible by the supervisor and me. The informed consent form with signatures will not be part of the report or any other deliverables, nor will they be stored/shared digitally in any way. All informed consent forms with signatures, and the filled questionnaires will be destroyed after the project is delivered and the grade received, at the latest during August 2017.

Voluntary consent
Your participation in the experiment is entirely voluntary, and you have the right to withdraw from this at any point without stating any reasons. If you decide to withdraw, all your gathered information will be deleted in front of you and not used in this study project. If you have any questions regarding the experiment, you can contact me (Viveka-93992176) or my supervisor (Frode- 93227262).

I have read the informed consent form and agreed to participate in the experiment.

.....
(Signature by participant, date)

Figure B.0.1: Informed consent form.

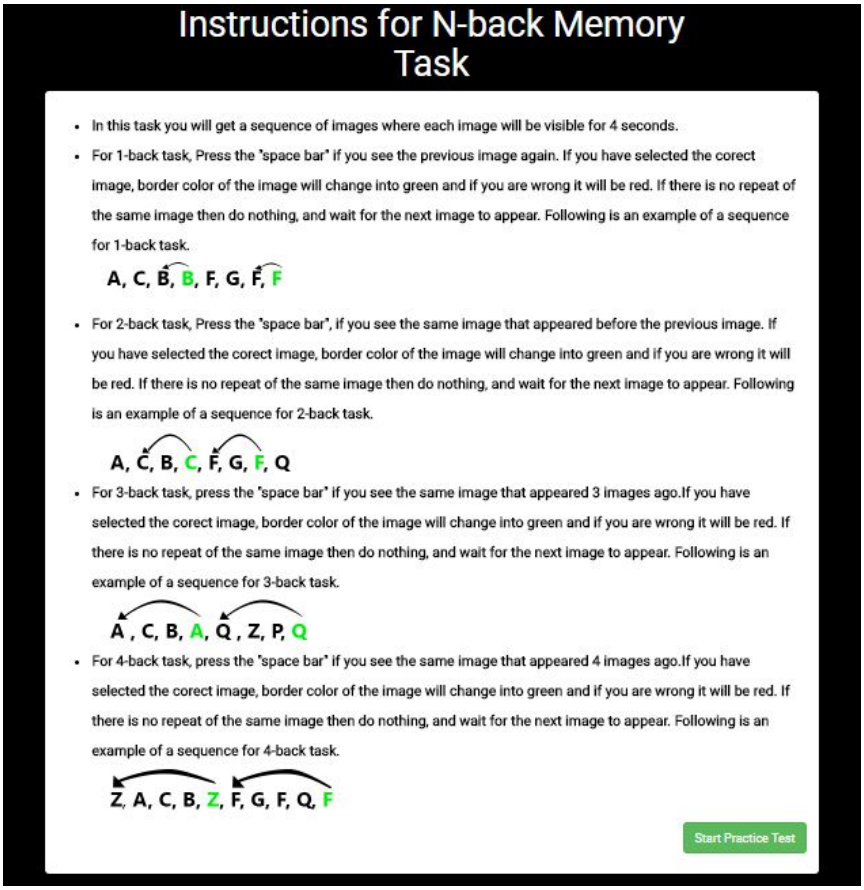

Figure B.0.2: Instructions for the participants.

C SPSS Analysis Settings

C.1 Paired-Samples t-test

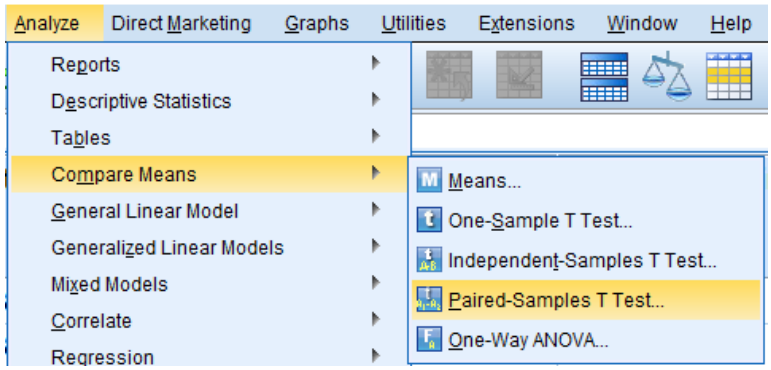

Figure C.1.1: Paired-Samples t-test - Step 1.

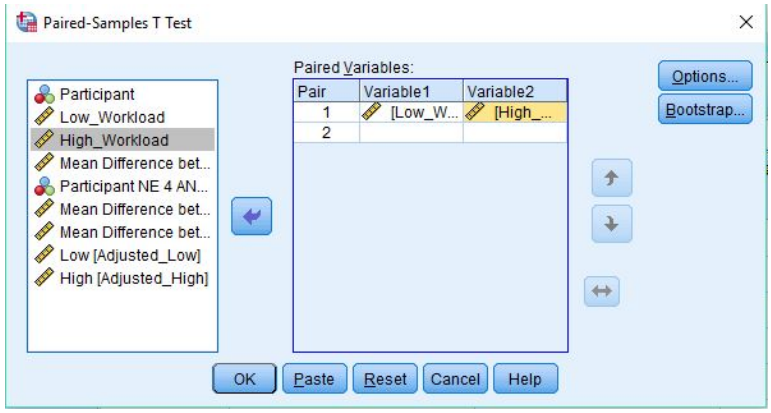

Figure C.1.2: Paired-Samples t-test - Step 2.

C.2 Repeated-Measures ANOVA

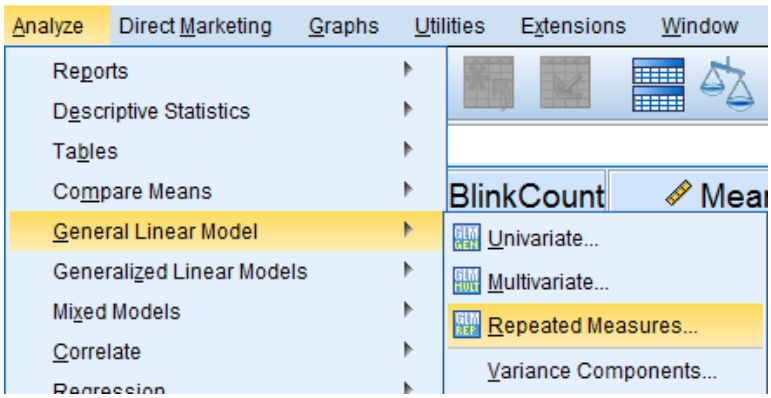

Figure C.2.3: ANOVA- Step 1.

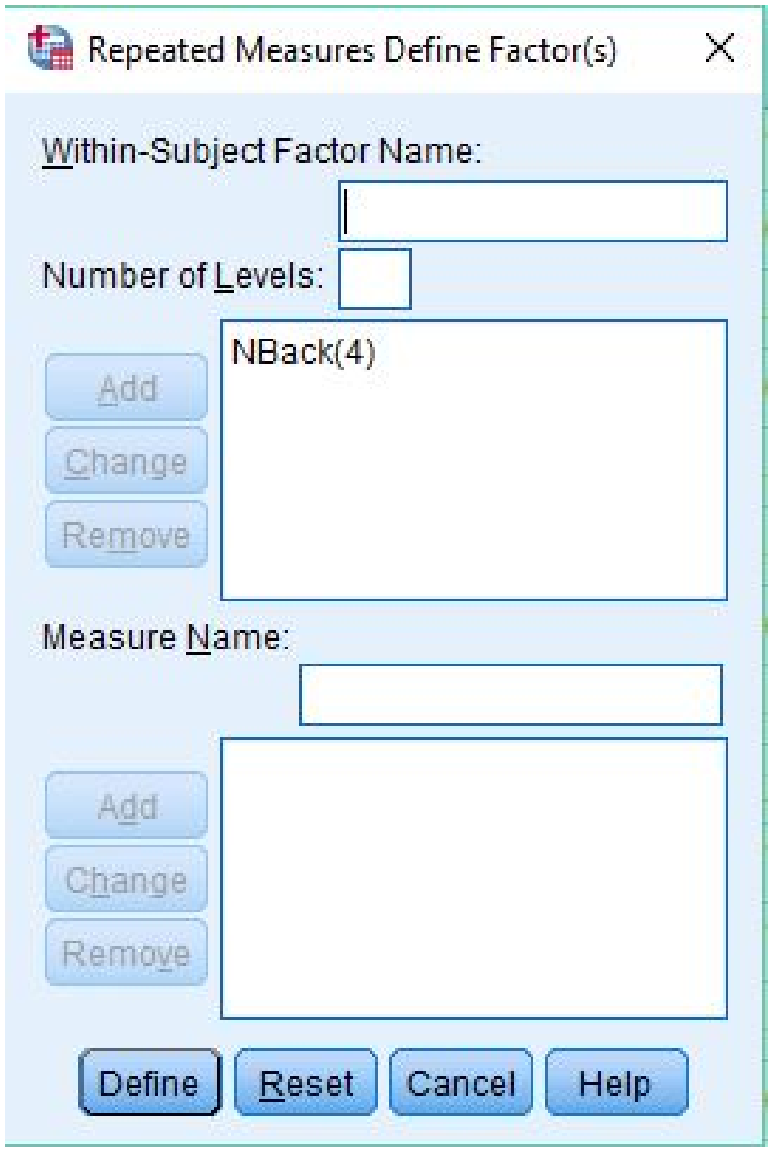

Figure C.2.4: ANOVA- Step 2.

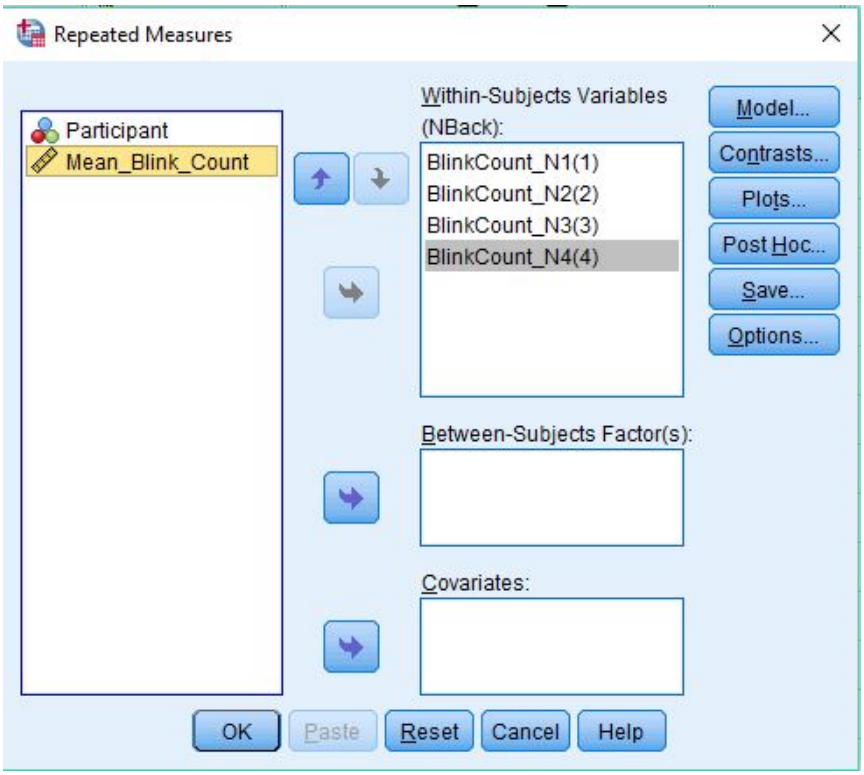

Figure C.2.5: ANOVA- Step 3.

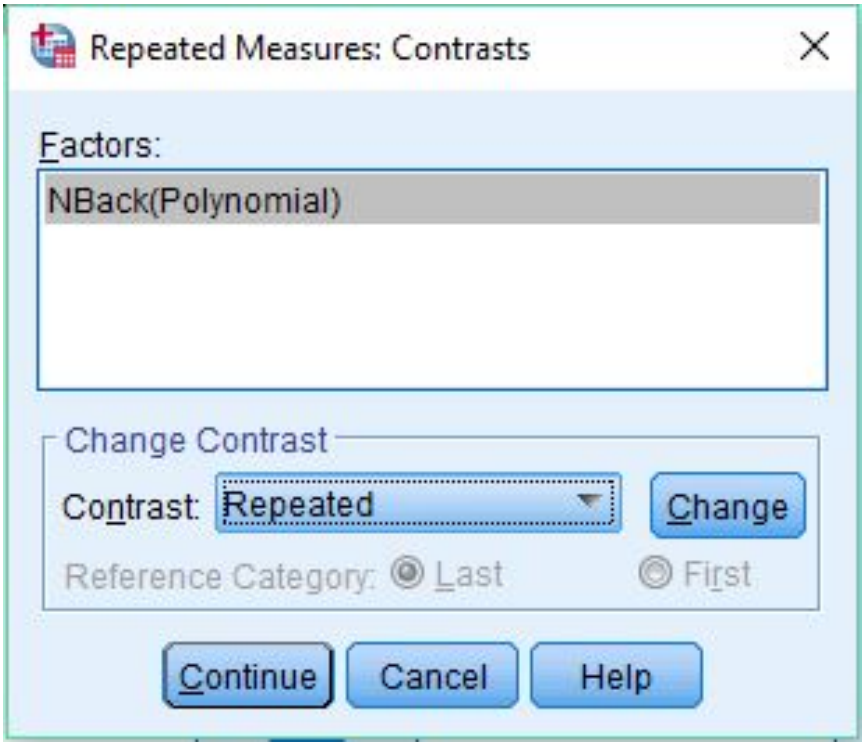

Figure C.2.6: ANOVA- Step 4.

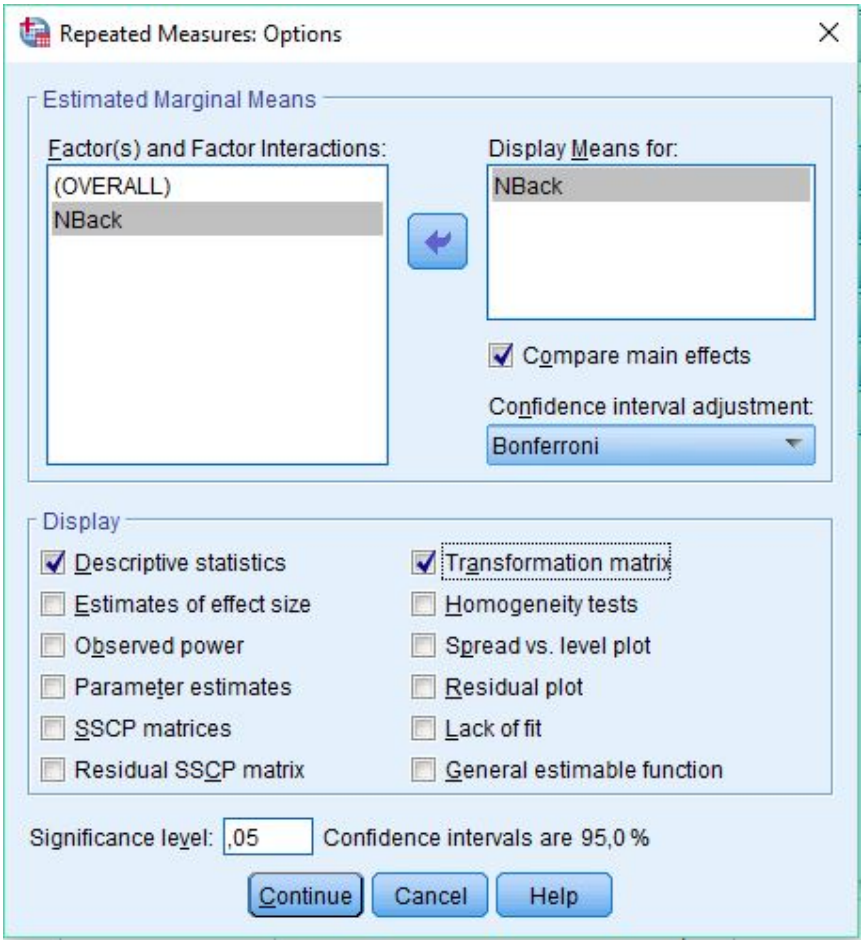

Figure C.2.7: ANOVA- Step 5.