EXAMENSARBETE INOM TEKNIK OCH HÄLSA,
AVANCERAD NIVÅ, 15 HP
STOCKHOLM, SVERIGE 2021

The criterion-related validity of pupil diameter as a measurement of mental fatigue in air traffic controllers

SANDRA JONASSON

KTH
SKOLAN FÖR KEMI, BIOTEKNOLOGI OCH HÄLSA

Abstract

Air traffic controllers face serious responsibilities and demanding work tasks on duty, which can lead to them experiencing mental fatigue. Due to the adverse effects that fatigue has on air traffic controllers' performance, it is important to be able to measure their fatigue levels while on duty in order to be able to intervene when the fatigue reaches critical levels. Pupil diameter is one possible mental fatigue measurement, but there are very few studies that have assessed its validity for measuring fatigue in air traffic controllers in the field. This study therefore aimed to assess the criterion-related validity of pupil diameter as a measure of fatigue in air traffic controllers on duty. Reaction time as measured by the psychomotor vigilance test was used as the gold standard fatigue measurement, while controlling for the effect of illumination on pupil diameter. A two-stage hierarchical multiple linear regression analysis was conducted to evaluate the prediction of pupil diameter from reaction time and illuminance. The result of this study indicates that neither RT nor illumination were significant predictors of the pupil diameter of air traffic controllers on duty. Overall, the result of this study indicates that a more complicated analysis with more data points on both illumination and pupil diameter might be necessary to be able to use illumination as a statistically significant predictor of pupil diameter in the field, and that the criterion-related validity of mean pupil diameter as a measurement of mental fatigue in air traffic controllers on duty may be low.

Acknowledgement

I would like to express my gratitude to Maximilian Peukert and Lothar Meyer at LFV for providing me with the opportunity to be a part of their interesting research, and for supervising me throughout the work of this thesis. I would also like to extend my special thanks to my primary supervisor Mats Ericson at KTH, who helped me finalize this thesis through valuable feedback and advice.

Table of Contents

Figures and tables

List of Abbreviations

LFV	Luftfartsverket
ATC	Air traffic control
ATCO	Air traffic controller
ATCOs	Air traffic controllers
EANS	Estonian Air Navigation Services
EEG	Electroencephalogram
ICAO	International Civil Aviation Organization
KSS	Karolinska Sleepiness Scale
Log10	Log base 10
mm	Millimeter(s)
ms	Millisecond(s)
min	minute(s)
PERCLOS	Percentage of eyelid closure over the pupil over time
PVT	Psychomotor Vigilance Test
RT	Reaction time

1. Introduction

Air traffic controllers (ATCOs) face serious responsibilities and demanding working conditions on duty, including high workload, stress, and complex work tasks (Chen et al., 2019). Their main work task consists of monitoring the air traffic to prevent aircraft collisions, both in and outside of the manoeuvring area, with the goal of optimizing the traffic flow and safety (ICAO, 2016; Svensson, 2020). In practice, this means that ATCOs are working with detecting and solving potential conflicts between flight paths to keep the air traffic safe and efficient (Svensson, 2020; Tello et al., 2018). In addition to the work tasks of ATCOs being demanding and complex, they also tend to be monotonous (Straussberger & Schaefer, 2007). Monotony in air traffic control (ATC) can appear both as a result of ATCOs having small amounts of traffic to manage in certain time periods such as during night shifts, or due to the fact that the traffic flow has some traits that are considered to be repetitive (Straussberger & Schaefer, 2007).

As the amount of air traffic that needs to be controlled and monitored is large and increasing, there also is a high and growing demand on ATC (Cebola & Kilner, 2010; Chen et al., 2019; Langan-Fox et al., 2009). This high ATC demand is expected to further increase the demands of ATCOs (Cebola & Kilner, 2010). As there are limits to how much traffic a human ATCO can work with, the airspace is divided into different sections (Svensson, 2020; Tello et al., 2018). On duty, one ATCO is then responsible for overlooking the air traffic in one such airspace sector. As air transportation moreover is available 24 hours a day, ATCOs are working in shifts both day and night to enable constant monitoring of the traffic (Cebola & Kilner, 2010; Chen et al., 2019). In addition to working day and night shifts, many ATCOs have different time schedules for their day and night shifts (Chen et al., 2019).

Because of the high demand, heavy workload, stress, monotony, complex work tasks, shift work and nightwork in the work of ATCOs, it is often considered as one of the most taxing and arduous jobs in the world (Chen et al., 2019). These challenging aspects of the work lead to ATCOs often experiencing fatigue on duty (Chen et al., 2019), which is a serious issue as it may have adverse effects on their performance (Bendak & Rashid, 2020; Zhang et al., 2019). When ATCOs are fatigued, their reaction time and stability decline, mistakes appear more easily, and their level of situation awareness decrease (Zhang et al., 2019). As ATC is part of the large safety critical system of air traffic, it is important that the controllers have a high level of performance and alertness to maintain a high level of air traffic safety.

The present study is conducted on behalf of Sweden's leading provider of ATC, Luftfartsverket (LFV). Due to the effects that fatigue has on ATCO performance and air traffic safety, LFV are interested in the possibility to measure fatigue levels in ATCOs while on duty. This in order to be able to interfere when the fatigue levels of ATCOs reach critical levels. LFV have identified ocular metrics gathered with eye tracking devices as a promising method for this purpose and are therefore interested in investigating what eye tracking parameters are indicators of fatigue in ATCOs.

There is a large number of studies that have studied eye tracking parameters as a possible measure of fatigue (Dinges & Grace, 1998; Hirshkowitz & Sharafkhaneh, 2017; McIntire et al., 2014a; McIntire et al., 2014b; Schleicher et al., 2008; Soares et al., 2013; Wilkinson et al., 2013; Yamada & Kobayashi, 2018). However, very few field studies have been conducted on eye tracking parameters for detection of fatigue in ATCOs specifically. In one study, Kuo et al., (2017) investigated whether the operators' gaze and pupil diameter could be used to monitor the level of fatigue and workload in ATCOs. The results of the study showed that the standard deviation of gaze could predict the workload of ATCOs, and that pupil diameter had a significant positive relationship with the level of fatigue in the controllers.

Because there is a very limited amount of field studies that have been conducted on the topic, the validity of eye tracking parameters for detecting fatigue in ATCOs in their natural working environment is not determined. The present study aims to assess the criterion-related validity of pupil diameter for detecting fatigue in ATCOs in the field and is therefore filling a gap in the existing literature on the topic.

2. Theoretical background

In the following section, the theoretical background of the present study will be presented. This includes descriptions of the concepts of vigilance and fatigue, what methods that are common for measuring fatigue, and what criterion-related validity is. The aim, purpose and delimitations of this study is also described in this section.

2.1 Vigilance and fatigue

Vigilance is a term used to describe the ability to sustain attention and to remain alert to stimuli over long periods of time (Helander, 2005; Warm et al., 2008). The systematic studying of vigilance begun during World War II, when it was discovered that radar operators missed certain weak signals on the displays that they monitored, especially in the end of the monitoring sessions (Warm et al., 2008).

Norman Mackworth was one of the pioneers in vigilance research (Warm et al., 2008; Hancock et al., 2016). During World War II, Mackworth conducted many experimental simulator studies on radar monitoring. The studies conducted by Mackworth showed that after only 30 minutes of monitoring, the accuracy of the operators' signal detection declined by about 10-15 percent. Mackworth's findings on vigilance have since been replicated several times. This decrease in vigilance levels is today referred to as *vigilance decrement,* which is a term that is defined as the gradual degradation of the ability to differentiate the occurrence of critical signals over time (Hancock et al., 2016). The phenomenon of vigilance decrement is often discussed in relation to monitoring operators that have to stay alert to detect different signals over long periods of time, and how to improve on the human-technology interaction in these situations (Hancock et al., 2016).

The concept of vigilance decrement is very closely related to the concept of *fatigue*. This is especially the case for the concept of mental fatigue, which in the literature sometimes is described to be the very same phenomenon as vigilance decrement (Haubert et al., 2018; Oken et al., 2006). Put simply, a decreasing level of alertness is considered to be highly related to an increase in the level of mental fatigue.

Fatigue is a term that is used in a broad manner to explain and describe sensations of sleepiness, tiredness, or physical exhaustion (Straussberger & Schaefer, 2007). While everyone knows what it feels like to be tired or fatigued, there is no rigorous definition of the phenomenon, and the different definitions in the literature vary to a great extent. One way of defining fatigue is that it is the likelihood of falling asleep, but there is a lot more to fatigue than being sleepy which is not captured in this definition (Matthews et al., 2012). One thing that for instance is not captured in this definition is that fatigue can take many different forms. There is *local fatigue* that is about a performance reduction of one organ, *central fatigue* that is when the whole organism is fatigued, *physical fatigue* or peripheral fatigue, as well as *mental fatigue* or central fatigue (Straussberger & Schaefer, 2007).

Another definition of fatigue, that has become popular and widespread within the aviation industry, is a definition that was made by the International Civil Aviation Organization (ICAO, 2016). They describe fatigue in the following manner:

> "A physiological state of reduced mental or physical performance capability resulting from sleep loss or extended wakefulness, circadian phase, or workload (mental and/or physical activity) that can impair a crew member's alertness and ability to safely operate an aircraft or perform safety-related duties" (ICAO, 2016, p. xii).

In this definition, it is apparent that there are many different sources of fatigue, and that fatigue is a physiological state that affects one's performance. In the present report, this definition will be used as a primary definition of fatigue due to the popularity of it in the aviation industry. In the following, sources of fatigue and its effects on human performance will therefore be elaborated on.

2.1.1 Sources of fatigue

There are many different sources of fatigue. Some of the most common sources of fatigue are related to sleep. One clear sleep related source is sleep deprivation i.e., when the duration of sleep is insufficient or the quality of sleep is poor (Matthews et al, 2012; Seedhouse et al., 2020). Most human adults need approximately seven to eight hours of good quality sleep per every 24-hour period, and even though most people prioritize to sleep for seven to eight hours a day, there is still a large amount of people who do not get enough sleep or good enough quality of their sleep, which then leads to fatigue (Seedhouse et al., 2020).

Another sleep related source of fatigue is interruptions to the 24-hour circadian cycle in sleepiness and wakefulness. The circadian cycle or rhythm can be described as the body's internal clock, with a biological tendency to be sleepy during the night and alert during the day (Caldwell, 2005). In other words, we are biologically programmed to become fatigued during the night and when there are interruptions to our circadian rhythm.

Fatigue can also be caused by carrying out various tasks and activities. When performing tasks that have a high degree of workload (mental, physical, or both), one tends to get fatigued (Federal Aviation Administration, 2013; Dasari et al., 2017; Triyanti et al., 2020). In the literature, this is commonly referred to as task-induced fatigue. In addition to high workload, the amount of time that is spent on a task is also known to be related to fatigue, which is referred to as "time-on-task" effects, which is an effect that has been widely studied (Nealley & Gawron, 2015; Trutschel et al., 2011).

Other tasks that are known to be inducing fatigue are tasks that are monotonous or boring (Straussberger & Schaefer, 2007). In fact, the state of monotony is so closely related to the state of fatigue that Straussberger and Schaefer (2007) in their article discussed the definitions of the two terms regarding whether or not they are the same phenomenon or not. According to Straussberger and Schaefer (2007), the two states of fatigue and monotony are similar in many ways, and they both lead to a disinterest in work. Monotony is caused by performing repetitive tasks and can therefore be avoided by introducing variation in the tasks or increasing the pace of work. On the other hand, fatigue is about a decrementing performance that is accelerated by increasing the pace of work.

Sleep deprivation, interruptions to the circadian rhythm and performing various tasks are three of the most important sources of fatigue (Matthews et al., 2012). There are however other sources of fatigue as well, including various health factors such as infections, drug use or nutritional factors (Matthews et al., 2012). These sources will however not be discussed in detail in this report.

2.1.2 Effects on ATCOs' performance

According to ICAO's definition on fatigue mentioned above, one of the most important aspects of fatigue is that it is a state of reduced physical or mental performance (ICAO, 2016). In this section, reduced performance induced by fatigue will be discussed. Since the focus of this work is fatigue in ATCOs, the focus here will be on reduced mental performance rather than physical performance, because the work tasks of ATCOs are demanding more mental effort than physical effort.

Whether fatigue is caused by lack of sleep, disruptions in circadian rhythms or extended time on task, fatigue makes the cognitive processes degrade to the extent where it can cause catastrophic breakdowns in performance (Gunzelmann & Gluck, 2009). By just staying awake and keep performing a work task for 18.5 - 21 hours, performance changes similar to those seen when having a 0.05 – 0.08% blood alcohol concentration can be observed (Arnedt et al., 2001).

Mental fatigue is known to result in several different impairments of one's *executive control* (van der Linden, 2010), which is a broad term that refers to cognitive functions that are required for being able to control one's thoughts, emotions, and actions (Brocki, 2007). Some effects that mental fatigue have been shown to have on executive control is difficulties in attentional focusing, information processing, and reduced planning and flexibility (van der Linden, 2010). In addition to these adverse effects on cognition, fatigue is also known to decrease the ability of decision making, memory, judgements, reaction time and situational awareness (Bendak & Rashid, 2020).

For ATCOs specifically, fatigue has a large impact on their performance. With the theoretical background on fatigue above, along with changing work hours and long work shifts, it is fair to assume that ATCOs have a high risk of experiencing both vigilance decrement and mental fatigue at work. Zhang et al. (2019) conducted a study on this topic, where they investigated how fatigue and stress are influencing the performance of ATCOs. They found that when fatigued, the controllers over all are both slower and lazier. They have longer reaction times in responding to pilots' calls and their reaction stability declines with fatigue. Moreover, they found that it is more likely that mistakes will occur in the radiotelephone communication when the controllers are fatigued.

2.2 Measuring fatigue

Researchers and clinicians have been grappling with the issue of lacking a clear definition of fatigue for more than 150 years (DeLuca, 2005; Hancock et al., 2012). As the meaning of the term "fatigue" tend to vary depending on in what situation or circumstance it is used, there are some issues with measuring fatigue that follow. One such issue is a paradox that has come to be known as "Muscio's paradox" (Hancock et al., 2012). This paradox originates from a foundational paper, in which Muscio (1921) discussed fatigue tests and whether it is possible to measure fatigue or not. In the paper, Muscio (1921) argued that if we do not have an unambiguous and rigid definition of fatigue, we cannot know with certainty that fatigue tests are measuring fatigue or something else entirely. 100 years later this paradox remains, as a definition of the term still not has been concluded on.

One solution that Muscio (1921) argued would solve these measurement issues was to simply not measure fatigue at all. However, this has largely been ignored and many efforts have been made to measure fatigue despite the terminological concerns (Aaronson et al., 1999). One reason for this is that it is hard to ignore fatigue as a phenomenon entirely, due to the important adverse effects that is has on human performance, both in general and in safety critical systems such as ATC specifically (Matthews et al., 2012). Today, fatigue is being measured to a large extent. It is often measured either as a subjective feeling or experience, or a direct or indirect performance decrement (DeLuca, 2005), but recent advances in for example neuroscience have also allowed for more physiological measurements (Matthews et al., 2012).

2.2.1 Performance-based measures of fatigue

One way of measuring fatigue is by using performance-based methods, which is done by carrying out measurements of the participants' task performance (Bendak & Rashid, 2020). This can be done in two different ways. One way is to measure the task performance of a task that is related to an operator's main task functions such as measuring how well a driver is steering a car when tired. One example of a study that did this in an ATC context is a study conducted by Zhang et al. (2019), where they measured how many mistakes that occur in ATC radiotelephone communication when the participants were fatigued.

The second way of measuring fatigue with performance-based methods is by measuring the participants' task performance of a secondary task, such as reaction time (Bendak & Rashid, 2020). The underpinning principle here is that fatigue is associated with functional impairment that makes response times slower as well as incorrect and/or failures to respond more frequent (Zuraida & Iridiastadi, 2011). One method that can be used to measure fatigue in this manner is the Psychomotor Vigilance Test (PVT).

2.2.1.1 Psychomotor vigilance test (PVT)

PVT is a fatigue test where the reaction time of participants are measured. It is widely considered to be the gold standard of fatigue measurements and has been validated multiple times since the original test was developed in 1985 (Arsintescu et al., 2017; Bendak & Rashid, 2020; Zuraida & Iridiastadi, 2011). It is also the most common metric to use for evaluating levels of vigilance and alertness, especially in studies that investigate sleep deprivation (Kilgore, 2010; Matthews et al., 2012).

The PVT is a simple test where the participant is pressing a button in response to visual cues that repeatedly appear on a screen (Arsintescu et al., 2017; Kilgore, 2010). The stimulus is presented at pseudo-random intervals that are ranging from 2 to 10 seconds. When the stimulus is presented, the participant is to press a button as quickly as possible in order to keep the reaction time as low as possible. Apart from measuring reaction time, PTV also allows for measuring how many times the button was not pressed when the visual cues did appear, as well as how many times the button was pressed when the visual cue did not appear (Arsintescu et al., 2017). The traditional task duration for the PVT was 10 minutes, but different versions have been developed of the test to suit different kinds of studies and study designs.

Some of the reasons that PVT is the most common and the gold-standard of fatigue measurements is that it is very sensitive to slowing of reaction times and the attentional lapses that are commonly occurring as a result of fatigue and/or sleep deprivation (Kilgore, 2010). Moreover, it is a highly reliable test that has been validated multiple times, and it shows very small learning effects (Kilgore, 2010).

3.2.2 Subjective measures of fatigue

Subjective metrics is another common and well-established way of measuring fatigue, that usually are both quick and easy to use (Golz et al., 2010). There are many different kinds of subjective scales for rating fatigue (Bendak & Rashid, 2020), and they can be structured either as questionnaires or subjective scales. Two of the most common subjective measures of fatigue is the Karolinska Sleepiness Scale (KSS) and the Samn-Perelli Checklist.

KSS is a validated self-rating scale of fatigue that was first developed by Åkerstedt and Gillberg (1990). The self-rating scale is simply used by presenting the participant with a scale with nine different levels of fatigue. The participant then determines which of the levels of fatigue that corresponds to their experienced level of fatigue. The original KSS ranged from 1 to 9, of which the odd scores contained written descriptions of the level of fatigue (Åkerstedt & Gillberg, 1990). The scale was later developed, so that it today includes written descriptions for all of the ten levels of fatigue (Golz et al., 2010). The levels are structured in the following manner: 1 completely alert, 2 very alert, 3 alert, 4 rather alert, 5 neither alert nor sleepy, 6 some signs of sleepiness, 7 sleepy (but no effort to keep awake), 8 sleepy (some effort to keep awake), 9 very sleepy (great effort to keep awake (fighting sleep) (Golz et al., 2010; Kaida et al., 2006).

Another subjective fatigue scale that is renowned and validated is the Samn-Perelli Checklist, which was developed for the purpose of estimating aircrew fatigue (Bendak & Rashid, 2020; Samn & Perelli, 1982). The Samn-Perelli Checklist is a 7-scale checklist of fatigue that works in a similar way as the KSS, where the participant is using the scale to determine which of the fatigue level matches their subjective experience of fatigue.

While both the Samn-Perelli Checklist and KSS are renowned measures of fatigue, there are some questions that have been raised regarding how reliable subjective measures like these are (Bendak & Rashid, 2020; Matthews et al., 2012; Tremaine et al., 2010). It seems like objective measures and subjective measures of fatigue and sleepiness do not always correspond (Tremaine et al., 2010), which one should be aware and mindful of when using subjective methods.

2.2.3 Physiological measures of fatigue

One common and objective way of measuring fatigue is by using physiological measures that monitor the physiological behaviour of the participants (Bendak & Rashid, 2020). In the beginning of the twentieth century, researchers started using various psychophysiological measures for measuring effects of sleep and time of the day, which was followed by the first EEG recordings in 1924 (Matthews et al., 2012). Today, EEG has become one of the most popular and widely used measures in fatigue related research on both sleep deprivation and sustained performance (Craig & Tran, 2012). As the brain activity is changing when a person becomes fatigued or drowsy, EEG is a suitable tool to use as it is measuring these said changes in brain activity (Craig & Tran, 2012). Moreover, EEG seems to precede performance impairments that are induced by fatigue, which means that it also can be used to predict and warn of fatigue (Matthews et al., 2012).

Other physiological measures of fatigue include measures of biochemical changes in the body, as well as measures of the cerebral blood flow (Matthews et al., 2012). Yet another way of measuring fatigue is by using eye tracking parameters such as blinking, pupil diameter and other eye movements, which is a method that has grown in popularity (Sikander & Anwar, 2018).

2.2.3.1 Eye tracking parameters

Eye tracking is a research tool for measuring different kinds of eye movements in subjects (Holmqvist & Andersson, 2017). It is a tool that is growing in popularity in a wide variety of disciplines such as cognitive psychology, psycholinguistics, electrical engineering, and usability analysis (Holmqvist & Andersson, 2017). Researchers in these and many other disciplines have found eye tracking parameters to be interesting for different purposes and research questions. In research on mental fatigue and vigilance, there are some parameters that have been of particular interest to researchers. These include parameters related to the subjects blinking, pupil diameter, saccades, and fixations.

Eye blinking parameters are some of the most commonly used eye tracking parameters for measuring fatigue and vigilance. Both the eye blink rate and the percentage of eyelid closure over the pupil over time (PERCLOS) have been shown to detect levels of alertness and fatigue (Dinges & Grace, 1998; McIntire et al., 2014b; Schleicher et al., 2008; Wilkinson et al., 2013; Yamada & Kobayashi, 2018). One relationship between fatigue and eye blinks is that as fatigue increases, the blink frequency and duration tend to increase as well (McIntire et al., 2014b). In other words, the level of fatigue correlates with the duration and frequency of blinks. In the PERCLOS fatigue metric, the eye blinks are not measured directly. Instead, the percentage of eyelid closure in a given period of time is measured to show droops and slow eyelid closures (Dinges & Grace, 1998). Dinges and Grace (1998) showed that the PERCLOS metric is both reliable and valid for detecting fatigue and has therefore been used in several fatigue studies since.

In contrast to the promising results of the blink studies mentioned above, one study (Trutschel et al., 2011) suggested that one should be cautious when using PERCLOS as a fatigue measurement. The reason for this is that PERCOS was shown to be less sensitive to fatigue levels than some other measurements that were used in their study. The conclusion of the study by Trutschel et al. (2011) was therefore that isolated measures like PERCLOS ideally should be used together with other fatigue measurements, rather than on its own.

Other eye tracking parameters that have been shown to be able to detect levels of fatigue are saccades and eye fixations. Saccades are the fast movements of the eyes between fixating points, and the fixation duration is usually defined as the time between these movements when the eyes are fixated at one point (Schleicher et al., 2008). The saccadic speed has been shown to be a reliable indicator of fatigue, where the speed of the saccades decreases as the level of fatigue in subjects increases (Di Stasi et al., 2014). Fixation duration on the other hand, is related to cognitive processing in vigilant subjects but has not been shown to be able to detect fatigue (Schleicher et al., 2008).

Another eye tracking parameter that has been shown to be able to detect levels of fatigue and vigilance is the pupil size (diameter) and its dilations (Hirshkowitz & Sharafkhaneh, 2017; McIntire et al., 2014a; Soares et al., 2013). This is due to the pupil size changing depending on the level of the nervous system arousal. As subjects start to feel sleepy, the parasympathetic nervous system is activated which leads to the pupil diameter decreasing (Hirshkowitz & Sharafkhaneh, 2017). When fatigued, the pupil size moreover tends to be instable as it decreases and oscillates at both a lower frequency and a higher amplitude (Hirshkowitz & Sharafkhaneh, 2017; Soares et al., 2013).

From an anatomical point of view, the autonomic nervous system is divided into two different and opposite parts: the sympathetic nervous system and the parasympathetic nervous system (Fotiou et al., 2000; Noyes & Barber-Westin, 2017). The sympathetic nervous system is responsible for the fight-or-flight response and prepares the body for stress by among other things increase the heart rate and blood pressure (Noyes & Barber-Westin, 2017). The parasympathetic nervous system on the other hand, is responsible for the rest and digestion reaction of the body when relaxing, eating, or resting. Simply out, it is undoing the effects of the sympathetic nervous system when the stressful situation has passed (Noyes & Barber-Westin, 2017). The functional balance between the sympathetic nervous system and the parasympathetic nervous system also determines the change in pupil diameter (Fotiou et al., 2000). When the sympathetic nervous system is stimulated, the pupil is dilating and the pupil diameter is increasing, and when the parasympathetic nervous system is stimulated the pupils constrict and the pupil diameter is decreasing (Fotiou et al., 2000; Noyes & Barber-Westin, 2017).

Apart from pupil diameter being sensitive to fatigue and sleepiness, it is also sensitive to many other environmental effects. The most prominent environmental factor is arguably exposure to light. When subjects are in a dark environment, the pupil automatically dialates in order to allow for more light to enter the eye and thereby improving the vision of the individual (Hirshkowitz & Sharafkhaneh, 2017). This is a widely known and studied effect and many different formulas of the relationship between light and pupil diameter have been suggested throughout the years. Watson and Yellott (2012) decided to review the different formulas to create a unified formula of the relationship. The relationship between the pupil diameter (in mm) and lux is illustrated graphically in their report. In figure 1A and 1B below, their illustration is shown with some modifications. Figure 1A shows the relationship with a field diameter of 60 degrees and a binocular viewing, i.e., when both eyes view the same adapting light). Figure 1B shows the relationship with a field diameter of 10 degrees and a monocular viewing, i.e., when only one eye is exposed to the light.

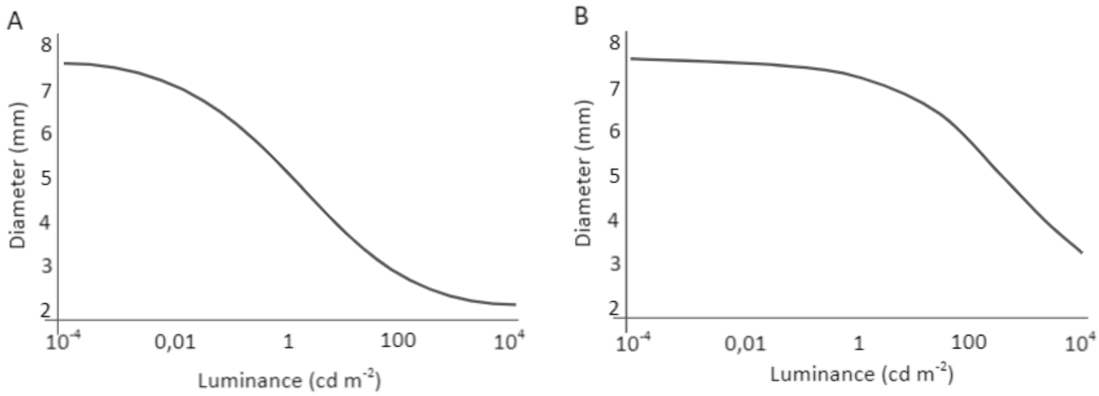

Figure 1. The relationship between light and pupil diameter. In figure 1A, the visual representation of the formula is shown with a field diameter of 60 degrees and a binocular viewing. In figure 1B, the visual representation of the formula is shown with a field diameter of 10 degrees and a binocular viewing.

Apart from illuminance and fatigue, there are many other factors that can influence the pupil diameter. Mental workload is one such factor, that has a small but significant effect on pupil diameter (Pfleging et al., 2016). When performing a difficult task with a high level of mental workload, the pupil diameter tends to increase, and when performing a task with a low degree of workload, the pupil diameter tends to decrease. Mental stress has a similar relationship with pupil diameter as mental workload does. As the level of mental stress increases, so does the pupil diameter (Yamanaka & Kawakami, 2009). This kind of increase in pupil diameter is based on the same principle as when the pupil diameter decreases due to fatigue, namely that the pupil diameter is affected by changes in the autonomic nervous system (Fotiou et al., 2000; Hirshkowitz & Sharafkhaneh, 2017; Noyes & Barber-Westin, 2017; Yamanaka & Kawakami, 2009).

2.3 Criterion-related validity

Before using a measurement in research on fatigue or any other phenomenon, it is important to know that the measuring method is actually measuring the phenomenon that it claims to do. When there is a strong connection between the concept of interest and its measurement, the measure method is said to be valid (DePoy & Gitlin, 2016). Put simply, the level of validity depends on how close the measurement method in question comes to representing the true definition of the researched phenomenon. Validity is an important factor to consider in research as it is measuring systematic errors. This means that if an instrument is not valid, there will be systematic bias in the measured data, which means the measurements also are invalid (DePoy & Gitlin, 2016).

One common type of validation is *criterion-related validity*. This type of validity is demonstrated by showing a relationship or correlation between the measurement that is to be validated and another instrument or method that has already been shown to accurately indicate the phenomenon or concept that is measured (DePoy & Gitlin, 2016; Miller & Lovler, 2016). In this type of validation, the new measurement method or instrument is usually put against the gold standard measurement, that is already known to be measuring what we want to measure with the new measurement. If there is a relationship or correlation between the results of the two measurement methods, the new measurement is said to be valid as it is capable of producing results that are related to the results of the valid gold standard measurement (DePoy & Gitlin, 2016; Miller & Lovler, 2016).

2.4 Aim and purpose

The purpose of this study was to assess the criterion-related validity of pupil diameter as a measure of fatigue in ATCOs on duty, by looking into whether their pupil diameter was related to their RT as measured by the fatigue gold standard measure PVT. The study also aimed to control for the effects of light on the pupil diameter by looking into if illumination could predict the pupil diameter in air traffic controllers on duty. The following research questions were formulated for this aim:

1. Does reaction time predict the pupil diameter in air traffic controllers on duty?
2. Does illumination predict the pupil diameter in air traffic controllers on duty?

2.4.1 Delimitations

Due to the limited time frame of the present thesis, some delimitations were made. In the present study, the pupil diameter is the only eye tracking parameter that is assessed regarding its criterion-related validity. This means that other eye tracking parameters such as the stability of the pupil size were excluded for validation in this study. Moreover, the light condition was the only variable that was controlled for in the study. This means that other factors that might have the potential to influence the pupil diameter like mental workload and stress were not included.

Another delimitation that was made for this study was to only focus on mental fatigue instead of central fatigue or any of the other types of fatigue discussed in *2.1 Vigilance and fatigue* in this report. This was done due to the nature of the work of ATCOs, which mainly is mentally demanding rather than physically demanding.

3. Method

In this section, the method of the present study will be presented. This includes descriptions of the study design, participants, materials, procedure, ethics, and data analysis.

3.1 Design

This study was conducted as a secondary research study, which means that the data that were analysed in this study originally were collected by someone else for another primary purpose (Johnston, 2014). The reason that the study was conducted in this manner was the outbreak of the COVID-19 pandemic, which complicated matters in conducting a primary research study. The original plan was to conduct a novel primary study on fatigue in ATCOs, but that study was cancelled due to the restrictions that the pandemic entailed.

The data that were used in this study were originally collected for a field study called Safe Tower Baseline (see Meyer, 2020) by researchers at LFV during the period of December 2019 to February 2020. The data collections took place in one Swedish ATC tower (ESNQ in Kiruna) and two different Estonian ATC towers (EETU in Tartu and EETN in Tallinn). The Safe Tower Baseline study had the purpose to "create an empiric baseline of air traffic controller work performance in a tower environment using performance shaping factors" (Meyer, 2020, p. 4). A large amount of data was therefore collected on performance shaping factors for ATCOs such as monotony, workload, stress, and fatigue in the field, in order to create a baseline for ATCOs' work performance.

Since there were many different kinds of quantitative data collected for the Safe Tower Baseline study, the types of data that were planned to be collected in the cancelled primary study mentioned above were already collected in the Safe Tower Baseline study. Moreover, the data that were collected in the Safe Tower Baseline study had not been explored from the view of the present study before. The data were therefore found to be suitable for answering the research questions of this study.

The design of the present study was a non-experimental regression study. This means that the variance in naturally occurring variables were investigated without any manipulation of the variables (Asamoah, 2014). The Safe Tower Baseline study that the data in this study originally were collected for was designed as a quantitative field study, which means that the quantitative data were gathered in a real and natural setting rather than in a laboratory (Aziz, 2017).

3.2 Participants

Eye tracking was measured for eight out of the twelve ATCOs that participated in the Safe Tower Baseline study. Four of the participants were wearing regular glasses and could therefore not wear the eye tracking glasses that were used when collecting the pupil diameter data. These four participants were therefore not included in the analysis of the present study. The eight participants whose data were included in the data analysis ($N = 8$) consisted of one woman and seven men (woman = 1 and men = 7). The mean age for this group was 31.5 years and their mean ATCO work experience was 7.7 years. In this category of participants, one of the participants were working in the Swedish ATC tower in Kiruna and seven of them were working in the Estonian airports in Tartu and Tallinn.

The participants were recruited by LVF and the Estonian Air Navigation Services (EANS). LFV recruited the participants who were working in the Swedish ATC tower in Kiruna by approaching the Chief of Operations at the airport, who then asked the ATCOs if they wanted to participate in the Safe Tower study. A similar recruiting method was used for the participants working in the Estonian ATC tower, where LFV’s contact person at EANS started a promotion for recruiting participants that wanted to participate in the Safe Tower study.

3.3 Materials

To gather the data that was analysed in this study, a PVT software program, a pair of eye tracking glasses and an illuminance measuring device were used as materials.

3.3.1 PVT software

The software program that was used to gather PVT data in the Safe Tower Baseline study and thus this study was the same that was used in the work of Najjar et al. (2014). The software program was running on a Microsoft Surface Go Tablet, with a Windows 10 operating system. In this software, each PVT test run lasted for three minutes. Within these three minutes, visual stimuli were presented at random times with intervals ranging from 3000 ms to 7000 ms. The stimuli consisted of an incrementing number that appeared on a screen that was connected to the computer on which the software was run. The reaction time was measured from the moment the number appeared on the screen to the moment when the participant pressed the spacebar key on the keyboard. The RT was then shown on the screen for 1000 ms to the participant.

3.3.2 Eye tracking device

The eye tracking device that was used for measuring the pupil diameter in this study was the Tobii Pro Glasses 2, that are developed in Sweden by Tobii (n.d.). This device is an eye tracking device that is worn by the participant as a regular pair of glasses. The glasses then measure the wearers eye movements with four different eye cameras (Tobii, n.d.). The glasses recorded many eye movement data such as pupil diameter (in mm), eye movement type (saccades, fixations), gaze position and pupil position.

3.3.3 Illuminance measuring device

A Konika Minolta CL-70F was used to measure the illumination in the ATC towers in the present study. Konika Minolta CL-70F is a CRI illuminance meter that is a lightweight and compact handheld instrument. It is measuring the both the colour and illuminance of light sources, which is shown in CRI, tristimulus values, illuminance, chromaticity, dominant wavelength, excitation purity, correlated colour temperature, and difference values from targets (Konica Minolta, n.d). In this study, the data that was recorded was the illuminance, dominant wavelength, and correlated colour temperature. The illuminance was measured in Lux (lx), which is a measure of illumination of a surface.

3.4 Procedure

During the Safe Tower Baseline field study, both PVT data, illuminance data and eye tracking data were collected throughout the ATCOs work shifts. See figure 2 below for a visual representation of an example of how the data was collected during one of the shifts of one of the participants in the study.

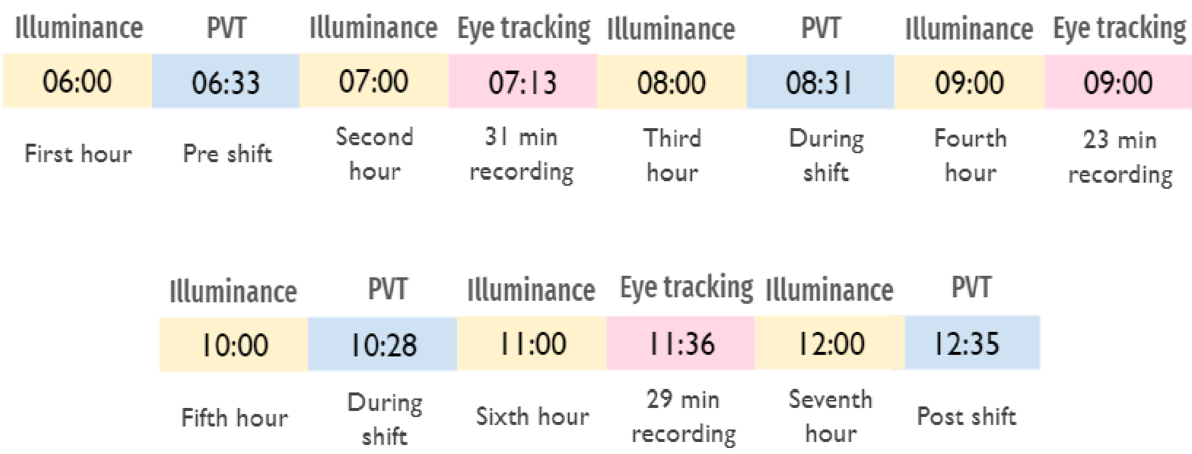

Figure 2. A visual representation of how the data collection was spread out during one of the shifts of a participant in the study.

PVT test runs were conducted every two hours, as well as when the work shift started, when leaving for and coming back from breaks, and at the end of the work shift. When a participant did the PVT test for the first time, one test run was conducted as a pilot test run to make them familiar with the test and the software before running the actual test. When it was time for a participant to do a PVT test, a researcher from LVF started the PVT software program and then handed over the control of the Microsoft Surface Go Tablet to the participant. The participants were instructed verbally based on a protocol/checklist. This included instructions to press the space key as fast as possible when the incrementing number appeared on the screen that was positioned before them. Each PVT test lasted for 3 minutes.

The illuminance was measured once every hour with the illuminance meter at the position of the head of the ATCOs when in working position. In some cases, when the ATCO was not present at the workstation, the measurements were done at the height of 120 cm above the floor, where the head of the ATCO would have been positioned if present. The measurements were carried out by measuring both in a horizontal direction and a vertical position. This means that two measurements were done with the Konika Minolta CL-70F every hour, one with the meter pointing to the ceiling and one with the meter pointing to the screens and windows of the working station.

The eye tracking recordings were not conducted as evenly distributed in time as the PVT tests and light measures were. Instead, the eye tracking recordings were recorded during eventful periods of the shifts. Ten minutes before the expected use of the runway, the recordings with the eye tracking glasses started, and they ended when they had left the control zone. When the runway was expected to be used, the researchers from LFV that were present during the work shift gave the eye tracking glasses to the participant. When the event was finished, the LFV researchers stopped the recording, and the participant took off the glasses.

3.5 Ethics

Throughout the work of this thesis, the Swedish Research Council’s guidelines on *Good research practice* (2017) were followed and taken into consideration. The data that was collected by researchers at LFV was done so with the consent of the participants, who were informed that their participation in the study was voluntary and that they could cancel their participation at any time. The data were handled according to GDPR, and a contract was set up between LFV and EANS for the Safe Tower Baseline study that contained legal considerations of handling of the data and GDPR, that was extended to cover the work period of the present study. Moreover, the Safe Tower Baseline study was permitted by the Estonian ethical scientific committee.

3.6 Data analysis

A hierarchical multiple regression was conducted as the data analysis of this study. Hierarchical regression is a sequential process where the predictor variables, or independent variables, are entered into the analysis in steps (Lewis, 2007). The order of the variable entry is based on theory and previous research, which allows for assessing the contribution of the predictors entered above and beyond the previously entered predictors as a way of statistically controlling for certain variables (Lewis, 2007). Before conducting the data analysis however, a number of things were done to prepare the data for the analysis. This included linear interpolation, calculating means, checking for assumptions and data transformations.

As mentioned in section *3.4 Procedure* of this report, the different measurements of pupil diameter and RT were spread out unevenly throughout the work shifts. To compensate for this, linear interpolation was done on the RT to replace missing values and to match the time of the eye tracking recordings. Linear interpolation is an imputation technique that can be used for replacing missing values in data points (Noor et al., 2015; Zhang & Chen, 2001). The technique assumes a linear relationship between the missing and non-missing values, which is illustrated in figure 3 below.

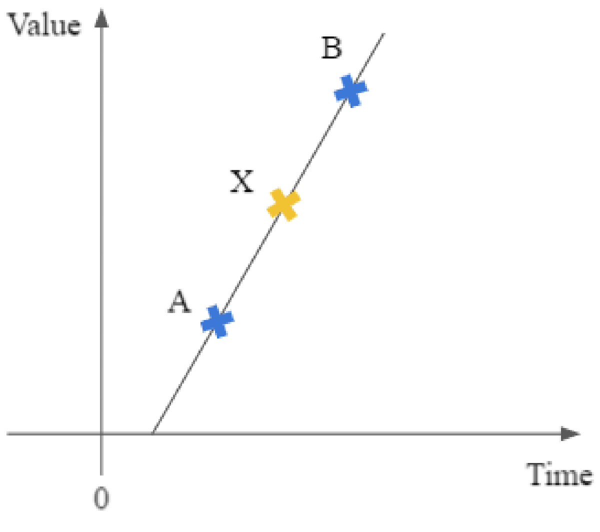

Figure 3. A visual representation of linear interpolation, where A and B are existing measured values, and X is a missing value which is calculated based on an assumption of linearity. This illustration is adapted from Zhang & Chen (2001) to describe the process of the linear interpolation done in this study.

The interpolation calculations were done with an interpolation tool developed by John D. Cook (n.d). The values that were used in the calculations were the mean RTs of the PTV tests that were done closest in time to the eye tracking recordings. Since the eye tracking recordings lasted for different number of minutes, the median time of the eye tracking recording in question was used to calculate the missing value of the RT. Put simply, the interpolated RT was calculated based on the two mean RTs measured by the two PTV tests closest to the eye tracking recording that it needed to match, and the median time of the eye tracking recording. The finished interpolated RT variable was therefore closer to the eye tracking recording in time than the original mean RTs were. The illumination values were not interpolated because the measures were done every hour, unlike the RT values that only were measured every other hour.

The pupil diameter variable was created by calculating the mean pupil diameter of each eye tracking recording that was available. In the original eye tracking data files, the pupil diameter was separated into left pupil diameter and right pupil diameter and contained thousands of data points. These were first calculated into means of both the left and right pupil diameter, which was then calculated into an overall mean for both of the eyes. This was done to create a variable that could be used when conducting the hierarchical multiple regression.

The illumination data did not need any preparation before conducting the statistical analysis. However, as mentioned in section *3.4 Procedure,* the illumination was measured both horizontally and vertically. For the hierarchical multiple regression, only one variable was going to be used. The horizontal measures were therefore chosen to be used as the variable in the regression.

In this stage of handling the data, the interpolated RT variable, the mean pupil diameter variable, and the horizontal illumination variable were ready for the statistical analysis. Here, time was considered as another variable that could have been accounted for due to the "time-on-task" effects and the human tendency to become fatigued the longer amount of time that is spent on a task (Nealley & Gawron, 2015; Trutschel et al., 2011). However, as the ATCOs that participated in this study had breaks where they also had the opportunity to rest and sleep, the time variable was not incorporated in the analysis.

The assumptions for conducting a hierarchical multiple regression according to Field (2013) were checked after the three variables were prepared for analysis. This was done to see if the model that the analysis resulted in would be generalizable i.e., working in other samples that the one that this data is from (Field, 2016). First the assumptions of linearity, normality, and homoscedasticity were checked. Normality was checked by creating histograms of all the variables and running a Shapiro-Wilks test, which is a test of normality. Then the linearity and homoscedasticity were checked by creating scatter plots with a fit line. The histograms, Shapiro-Wilks tests, and the scatter plots showed that the data did not meet the assumptions of linearity, normality, and homoscedasticity. The data had a positive skew and were not entirely linear or homoscedastic. One way of overcoming this problem is by transforming the data with the log base 10 (log10) (Field, 2013). Therefore, all the variables that were to be used in the analyses were transformed with log10.

When the variables of illumination, interpolated RT and mean pupil diameter were transformed with log10, the data were normally distributed, and the homoscedasticity of the date were largely improved. However, there were still no clear linear relationship between the interpolated RT variable and the mean pupil diameter variable. Moreover, the measures in this study were repeated measures, which violates the assumption of independence of linear regressions. The repeated measures between participants were not accounted for in the analysis. Despite these two violations of the assumption of hierarchical multiple regressions, the analysis was run on these data. Note that this has implications for the generalizability of the model (Field, 2016). This is an issue that will be further discussed in the section *5.2 Method discussion* in this report.

Before running the hierarchical multiple regression, a few more assumptions were checked. The multicollinearity was checked to assure the independent variables did not correlate with each other, which they did not ($R < 0.7$). A probability-probability plot was also plotted to check the overall linear relationship between the two transformed (log10) independent variables and the dependent variable, which was linear. Moreover, the residual statistics were checked to assure that the standard residual did not have values that exceeded -3 or 3, and that Cook's distance was not greater than 1, which they were not. Finally, the tolerance and VIF were checked. The tolerance was greater than .1 and the VIF was less than 10, which also meets the collinearity assumption of the analysis.

Once the data were prepared for analysis and checked for assumptions, the hierarchical multiple regression was run in IBM SPSS Statistics 26. In total, 39 cases with matched mean pupil diameter, PVT and illumination were used for the analysis. The regression was run with the method "forced entry" with pupil diameter (transformed with log10) as the dependent variable. The independent variable illumination (transformed with log10) was entered in the first stage to control for its effect on the dependent variable (Field, 2013). Then the second independent variable interpolated RT (transformed with log10) was entered in the second stage.

4. Result

To assess the criterion-related validity of pupil diameter as a measure of fatigue in ATCOs, a two-stage hierarchical multiple linear regression analysis was conducted to evaluate the prediction of pupil diameter from RT and illuminance. At stage one of the regression, the illuminance variable was entered to control for the effect that illuminance has on pupil diameter. The result of the first stage of the analysis showed a model not to be statistically significant ($F(1, 37) = .020$, $p = .157$), as the p-value was greater than the level of statistically significance that was set at .05. The R^2 value of this regression model was .053 which suggests that the illuminance variable accounts for 5.3% of the variance in the pupil diameter variable, and that the remaining 94.7% of the variance cannot be explained by the illuminance variable alone.

At stage two of the hierarchical multiple linear regression, the RT variable was added to the analysis. The result of the second stage of the analysis showed a model not to be statistically significant ($F(2, 36) = 0.14$, $p = .257$). The R^2 value of this regression model was .073 which suggests that the addition of the RT variable to the first stage model accounts for 7.3% of the variance in pupil diameter and that the remaining 92.7% of the variance cannot be explained by the illuminance and RT variables alone. A summary of all the regression statistics is shown in table 1 below.

Table 1.
Summary of the hierarchical multiple regression analysis for variables predicting pupil diameter (*N* = 8).

	B	SE B	β	*p*-value
Model 1				
Constant	.603	.044		<.001
Illuminance	-.036	.025	-.231	.157
Model 2				
Constant	1.903	1.501		.213
Illuminance	-.040	.025	-.260	.121
Interpolated RT	-.517	.596	-.142	.392

Comment. B = Non-standardized regression coefficient. Shows the relationship between pupil diameter and each independent variable, where a higher negative or positive value is indicating a higher negative or positive relationship. SE B = Standard error of B. Indicating how much the values of B would vary across different samples. β = Standardized regression coefficient / beta. Ranges from -1 and 1 where a value that is closer to either -1 or 1 shows a stronger relationship between the pupil diameter and the independent variable. *p*-value = the probability that the independent variable can predict the pupil diameter where a value less than .05 is significant.

4.1 Does illumination predict the pupil diameter in air traffic controllers on duty?

The hierarchical multiple linear regression showed that the illumination variable does not predict the mean pupil diameter in ATCOs on duty. The variable accounted for 5.3% of the variance in pupil diameter, and it is not a statistically significant proportion of variance. Therefore, the model does not have predictive utility.

The non-standardized regression coefficient (B in table 1 above) was -.040 for the illumination variable. This indicates that for each unit increase in the illumination variable, a corresponding -.040 unit increase would be predicted for the pupil diameter variable if the relationship was significant. However, the p-value of the corresponding t-test was .121 and illumination was therefore not considered to be a significant predictor of pupil diameter, or to have predictive utility.

4.2 Does reaction time predict the pupil diameter in air traffic controllers on duty?

The hierarchical multiple linear regression showed that RT does not predict the pupil diameter in ATCOs on duty. The stage two model with both illumination and RT as independent variables showed that the two variables together accounted for 7.3% of the variance in mean pupil diameter, and it is not a statistically significant proportion of variance. This means that the model does not have predictive utility.

The difference between the R^2 value for the first stage model and the second stage model was .019. This suggests that when controlling for the illumination variable, 1.9% of the variance in the dependent variable can be explained by the RT variable. Moreover, the partial unstandardized regression coefficient (B in table 1 above) was -.517 for the RT variable. This means that if the relationship were significant, for each unit increase in the RT variable, we would have predicted a corresponding -.517 unit increase in the pupil diameter variable after controlling for the illumination variable. The p-value of the corresponding t-test was .392 and RT was therefore not considered to be a significant predictor of pupil diameter, or to have predictive utility.

5. Discussion

The following sections consists of three main parts, a result discussion, a method discussion, and recommendations for future research. In the result discussion, the results of this study will be discussed in relation to the theory presented in *2. Theoretical background* of this report, as well as what implications they might have. In the method discussion, the method of this study will be critically discussed in relation to the result of this study. Finally, some recommendations for future research on the topic will be discussed.

5.1 Result discussion

To investigate the criterion-related validity of pupil diameter as a measurement of mental fatigue in ATCOs, a hierarchical multiple linear regression analysis was conducted on data that were gathered for the Safe Tower Baseline study conducted by Meyer (2020). RT that was measured using PVT was used as the criterion or gold standard measure of fatigue that were used to see whether it could predict the pupil diameter of ATCOs on duty, while also controlling for the effect of illumination on pupil diameter.

The result of this study indicates that neither RT nor illumination were significant predictors of the pupil diameter of ATCOs on duty. In the following sections, these results will be discussed in relation to literature and the method of this study.

5.1.1 Does illumination predict the pupil diameter in air traffic controllers on duty?

In the hierarchical multiple linear regression analysis, the illumination variable was used as a control variable because of the clear results from the multiple studies showing that there indeed is a strong relationship between illumination and pupil diameter (Hirshkowitz & Sharafkhaneh, 2017; Watson & Yellott, 2012). The result of the present study showed that illumination cannot predict the variance in pupil diameter, which is notable due to the previous research on the topic suggesting the opposite.

One possible reason for why the illumination could not predict the pupil diameter of ATCOs in this study might be because the pupil diameter variable consisted of the mean pupil diameter of the different eye tracking recordings. Most of the eye tracking recordings lasted approximately 20 minutes. The light conditions might therefore have been varying greatly during the time period when the eye tracking device was recording. There are many factors that affect the illumination levels in ATC towers such as blinds, curtains, tinted windows, screens, daylight, and artificial lights. If an ATCO decided to turn on or off an extra computer screen or light during an eye tracking recording, this could have created disturbances in the data because the illumination during that recording would not have been constant or consistent with the value that was measured horizontally with the illuminance meter that hour. In other words, while the mean pupil diameter was measured over a period of several minutes, the corresponding data point of illumination was measured once only, and they are therefore not matching entirely in terms of time. This type of dynamics is hard to control for in field studies in this type of environment, as the ATCOs' working conditions have to be considered while collecting the data.

The finding of not being able to predict mean pupil diameter with illumination is overall important, because it indicates that a measure that is as simple as the mean pupil diameter in eye tracking recordings likely not will be able to be predicted by illumination in future field studies on the topic either. Because the data that were used in this study were gathered in the field, it is hard to keep track of the illumination levels as they are not controlled for at all. As the data collection was not designed for the specific purpose of this study, the illumination was only measured once an hour. A possible way of measuring the illumination in future studies is to measure the illumination simultaneously with the eye tracking parameters. A more complicated analysis with more data points on both illumination and pupil diameter, might be necessary to be able to use illumination as a statistically significant predictor of pupil diameter. However, more studies need to be conducted on the topic to draw such conclusions.

5.1.2 Does reaction time predict the pupil diameter in air traffic controllers on duty?

In this study, RT that was measured by PVT was used as the gold standard fatigue measurement to assess the criterion-related validity of pupil diameter as a measurement of mental fatigue in ATCOs on duty. RT was not considered to be a significant predictor of pupil diameter or to have predictive utility in this study. This result was an unexpected result, as previous research that has been conducted on the topic has indicated that pupil diameter is able to detect levels of fatigue and vigilance decrements (Hirshkowitz & Sharafkhaneh, 2017; Kuo et al., 2017; McIntire et al., 2014a; Soares et al., 2013).

While PVT is the fatigue measurement that generally is considered to be the gold standard (Arsintescu et al., 2017; Bendak & Rashid, 2020; Zuraida & Iridiastadi, 2011), it may be argued that the participants in this study might not have experienced fatigue levels that were high enough for them to experience a significant decrement in RT performance. Basner & Dinges (2011) studied fatigue as a result of sleep deprivation and found that the highest effect sizes of mean RT measured by PVT were observed after being awake for 25 to 27 hours, while the lowest effect sizes were observed after being awake for 11 hours. During a normal work shift, ATCOs might experience a certain amount of fatigue due to the arduous nature of their work and working irregular hours (Chen et al., 2019), but it is not determined whether or not the fatigue ever reaches fatigue levels like the ones observed in Basner & Dinges (2011), where a drastic decrease in RT performance was found. There might also exist individual differences in RT performances among the participants to begin with, which would then require higher fatigue levels in order to be able to detect an overall pattern of RT performance decrements, as all of the participants were analysed as one group. However, as PVT indeed is sensitive to fatigue (Miller, 2012) and has been validated several times for the purpose (Arsintescu et al., 2017; Bendak & Rashid, 2020; Zuraida & Iridiastadi, 2011), the fatigue levels of the ATCOs that participated in this study should have been able to be detected correctly.

Another thing that might have affected the results of this research question was the uneven measuring of the RT and eye tracking recordings. As the eye tracking was recorded sporadically throughout the shifts of the ATCOs, and the PVT was measured every other hour, they were not measured simultaneously. The interpolation of the RT variable was done to correct for this. Interpolation assumes linearity and it is a method used for replacing missing values in data points (Noor et al., 2015; Zhang & Chen, 2001), but it is not certain that the fatigue and RT performance was following an exact linear pattern in between each eye tracking recordings for each of the participants. This likely affected the result of the present research question. In future studies that will not be conducted as secondary research studies, the different measures should be measured at the same time or very close in time in order to not have to account for missing values as it alters the variable.

As aforementioned, there are other factors than fatigue (and illumination) that might affect the pupil diameter, which potentially can explain the fact that only 1.9% of the variance in pupil diameter could be predicted by the fatigue measurement RT. Mental workload and stress are two such factors that effects the pupil diameter (Pfleging et al., 2016; Yamanaka & Kawakami, 2009). The relationship between pupil diameter and mental workload and stress is that when one is performing tasks that are stressful and have a high degree of mental workload, the pupil diameter tends to increase, and when performing tasks with a low degree of mental workload and stress, the pupil diameter tends to decrease (Pfleging et al., 2016). To compare this with the relationship between fatigue and pupil diameter, the pupil diameter tends to decrease as a result of fatigue (Hirshkowitz & Sharafkhaneh, 2017; McIntire et al., 2014a; Soares et al., 2013). Arguably, the mental workload and stress of the participants in this study might have been particularly large, as the work tasks of ATCOs can be both complex and demanding (Chen et al., 2019). This could in turn have led to an increase in the recorded pupil diameter. Especially considering that the pupil diameter only was measured during the periods when the runway was used and the ATCOs were actively working. The effect of mental workload and stress on pupil diameter is however not very large (Pfleging et al., 2016), so there probably exists other factors that might have been influencing the pupil diameter as well considering that only 1.9% of the variance in pupil diameter could be predicted by RT. As no measures of mental workload or stress were included in this study though, no conclusions can be drawn regarding their actual effect on pupil diameter.

Another aspect to note regarding the result that the RT variable only accounted for 1.9% of the variance in the pupil diameter variable is time-of the day effects. The 24-hour circadian cycle in sleepiness and wakefulness leads to a tendency of being fatigued during the night and alert during the day (Caldwell, 2005). The different data collections of this study were spread out during the day, and were done both during night shifts, morning shifts and evening shifts. As this was not accounted for in this study, the level of fatigue might have been varying in the different types of shifts. This would have led to greater variations in the ATCOs' fatigue levels which is a benefit when using the variable as a predictor of pupil diameter, because it acts as the gold standard that pupil diameter is compared to.

Overall, the result of this study indicates that the criterion-related validity of pupil diameter as a measurement of mental fatigue in ATCOs may be low. As the fatigue measurement RT is not a statistically significant predictor of the pupil diameter in ATCOs, it seems like other methods for measuring fatigue levels in ATCOs in field studies might generate more valid results. As discussed above, there are many other things that can affect the pupil diameter apart from the level of mental fatigue and illumination, and it might therefore be a measurement that is hard to get valid results from.

Some of the other fatigue measurements that have been discussed in the theoretical background of this report is subjective measurements like KSS and the Samn-Perelli Checklist, EEG, and other eye tracking parameters such as eye blinking, saccades, and fixations. The objective measures have an obvious advantage in fatigue studies in that they are not suffering from the subjectivity and individual variations that exists in subjective measures (Matthews et al., 2012). EEG is an objective measure that might not be affected by as many different factors as the pupil diameter, but it may also be hard to use EEG in fields studies where ATCOs need to be able to work without being interrupted by the measuring device. On the other hand, other eye tracking parameters such as eye blinking, saccades, and fixations are just as easy as pupil diameter to record in the field. Therefore, it would be interesting to investigate the validity of these measurements for the purpose of detecting mental fatigue in ATCOs on duty in future studies.

5.2 Method discussion

The present study was conducted as a secondary research study which means that the data was collected for another study with another purpose (Johnston, 2014). This study design had the implication that the measurements were done in a way that did not fit entirely with the purpose of this study. As discussed above, the eye tracking measurements, the illumination measurements, and the PVT measurements were not done simultaneously which did not benefit the analysis of this study. If a new data collection would have taken place as originally planned for, the measurements could have been done in a way that suited the study entirely and the linear interpolation that was done for the RT variable would not have been needed as the measurement times would already be matching.

The primary study that the present secondary research study was based on, Safe Tower Baseline study (Meyer, 2020), was a field study. This has many implications for the present study as a whole as there are many environmental factors in field studies that can affect both the data and results (Aziz, 2017). One key aspect of this is that the surrounding environment of the participants and the context cannot be controlled for in the same way as in experimental studies. As discussed above in *5.1 Result discussion*, there were many different factors in the ATC towers that could not be controlled for, which might make it harder to see relationships between different variables. In other words, the context and environment of this study was not controlled and optimized for observing changes in pupil diameters in the same way that a simulator study or any other kind of experimental study might have been. The uncontrollability of the context and environment in the field moreover makes it harder to determine the generalizability of the study, as it is hard to precisely characterize the field environment (Aziz, 2017).

Another challenging aspect of ATCO field data is that the quality of the eye tracking data can suffer (Li et al., 2020). Both the natural movements of controllers, illumination and reflection of visual displays can cause excess noise in the data. This is important as the quality of the data has an effect on detection of different eye tracking parameters.

A large advantage of data gathered in the field, however, is that it actually is generalizable to real-life contexts (Aziz, 2017). This is especially advantageous in the context of validation of fatigue measurements in ATCOs on duty, as the fatigue measure actually needs to be valid in the normal working conditions of ATC towers rather than in a controlled lab environment. Even though a fatigue measurement would be able to show valid results for measuring fatigue in ATCOs in a laboratory, this would still not mean that it would be able to produce valid results in the real work environment of an ATC tower. If a fatigue measurement, like as eye tracking parameters in this case, is supposed to be used for the purpose of alarming of critical fatigue levels of controllers on duty, it is most important that the measurement is valid for that specific purpose.

When checking the assumptions of the hierarchical multiple linear regression, no clear linear relationship between the interpolated RT variable and the mean pupil diameter variable was found. The variables that were used in the regression were transformed with log10 to improve on the linearity, but in spite of this there was no clear linear relationship between the two variables. Despite this, the RT variable was included in the regression analysis. This is a violation of the linearity assumption that is assumed for conducting a hierarchical multiple linear regression (Field, 2013; Field, 2016).

The repeated measures of this study were not accounted for and the assumption of independent measures for linear regressions was not met. Even though the measures were repeated between the participants, the regression analysis was run on the data. One method that could have accounted for the repeated measures would have been hierarchical linear modelling (Garson, 2013). This is a type of modelling that is used to analyse the variance in a dependent variable when the independent variables are at varying hierarchical levels (Woltman et al., 2012). I future studies on the topic of this study, the assumption of independent measures for linear regression should be taken into consideration, and hierarchical linear modelling is one option for doing so. Note that although the names of hierarchical regression (that was used in this study) and hierarchical linear modelling are similar, it is two distinct methods.

When all the assumptions are not met in a linear regression like in this study, the estimates of the *B*s of the model cannot be trusted. This has the implication that the model will not be generalizable (Field, 2016). In other words, it cannot be assumed that the model works in other samples than the one that the data was collected from. This means that the *B*s of the models that this study resulted in only works for the sample that was used in this study. No clear conclusions can therefore be drawn from this study for other samples due to the nature of the data that was used.

5.4 Recommendations for future research

As mentioned above, the regression models of this study are not generalizable to other samples. More studies are therefore needed to be able to draw conclusions on the validity of pupil diameter as a measurement of mental fatigue in ATCOs.

Future studies of the validity of pupil diameter as a measurement of mental fatigue in ATCOs should be conducted as primary research studies in order for the data collection to be planned with the research questions as a basis. This would likely allow for collection of data that are better suited for conducting statistical analyses like the hierarchical multiple linear regression that was done in this study. Moreover, such studies should take into consideration that there are many external factors that might influence the results of field studies. The effects of such factors could be minimized by keeping the environment as controlled as possible. For example, one ATC tower should be used for conducting the data collection to allow for a more consistent environment throughout the study, rather than using three different ATC towers like in this study.

It would also be interesting to investigate whether there exist any major differences between both the individuals and different groups of people in RT performance and pupil diameter. As mentioned above, there were many different conditions and different types of individuals that were a part in this study. Such analyses could help improve the understanding of which of the environmental factors were affecting the study and in what ways.

Finally, it would be interesting to study the validity of other eye tracking parameters for measuring fatigue in ATCOs on duty. In this study, eye blinking parameters, saccades, and fixations have been brought up as other possible fatigue measurements. These eye tracking parameters might not be as sensitive to illumination as pupil diameter and may therefore work better in the field. Before one or more of these parameters could be used as a fatigue measurement in the field however, it is important that they are validated to make sure that they produce valid results for the context.

6. Conclusion

In this study, a hierarchical multiple linear regression analysis was conducted to assess the criterion-related validity of pupil diameter as a measurement of mental fatigue in ATCOs on duty. Two research questions were formulated for this, which will be answered separately below. Note that both the linearity assumption and the independent measures assumption of the analysis was not met, which means that the conclusions here might not be true for other samples than the one that was used in this study (Field, 2016).

6.1 Does illumination predict the pupil diameter in air traffic controllers on duty?

In this study, the illumination variable was not a statistically significant predictor of the mean pupil diameter variable. The illumination variable only accounted for 5.3% of the variance in the pupil diameter variable.

This result was an unexpected result, as it contradicts the solid previous research on the topic. It is an important finding though, as it indicates that the mean pupil diameter of eye tracking recordings is a measure that is too simple to be able to be predicted by the illumination that was measured that same hour. A statistical analysis with more data points on both the illumination levels and pupil diameter might therefore be necessary to be able to use illumination as a statistically significant predictor of pupil diameter.

6.2 Does reaction time predict the pupil diameter in air traffic controllers on duty?

The RT variable was not a statistically significant predictor of the mean pupil diameter variable in this study. When the illumination variable was controlled for, only 1.9% of the variance in the pupil diameter variable could be explained by the RT variable. This result indicates that the criterion-related validity of pupil diameter as a measurement of mental fatigue in ATCOs on duty may be low for the sample used in this study. As there are many other factors than the illumination and fatigue levels that can influence the pupil diameter, there might exist other measurements that are more valid for the purpose. Some other eye tracking parameters that have been discussed as possible mental fatigue measurements in this report are eye blinks, saccades, and fixations. An interesting topic for future studies on the topic would therefore be to study the validity of these and other eye tracking parameters for measuring mental fatigue in ATCOs on duty.

7. References

Aaronson, L. S., Teel, C. S., Cassmeyer, V., Neuberger, G. B., Pallikkathayil, L., Pierce, J., Press, A. N., Williams, P. D., & Wingate, A. (1999). Defining and measuring fatigue. *Journal of Nursing Scholarship, 31*(1), 45–50. https://doi.org/10.1111/j.1547-5069.1999.tb00420.x

Arnedt, J. T., Wilde, G. J., Munt, P. W., & MacLean, A. W. (2001). How do prolonged wakefulness and alcohol compare in the decrements they produce on a simulated driving task?. *Accident Analysis & Prevention*, *33*(3), 337-344. https://doi.org/10.1016/S0001-4575(00)00047-6

Arsintescu, L., Kato, K. H., Cravalho, P. F., Feick, N. H., Stone, L. S., & Flynn-Evans, E. E. (2019). Validation of a touchscreen psychomotor vigilance task. *Accident Analysis and Prevention*, *126*, 173–176. https://doi.org/10.1016/j.aap.2017.11.041

Asamoah, M. K. (2014). Re-examination of the limitations associated with correlational research. *Journal of Educational Research and Reviews, 2*(4), 45-52.

Aziz, H. A. (2017). Comparison between field research and controlled laboratory research. *Archives of Clinical and Biomedical Research, 1*(2), 101-104.

Basner, M., & Dinges, D. F. (2011). Maximizing Sensitivity of the Psychomotor Vigilance Test (PVT) to Sleep Loss. *Sleep, 34*(5), 581–591. https://doi.org/10.1093/sleep/34.5.581

Bendak, S., & Rashid, H. S. J. (2020). Fatigue in aviation: A systematic review of the literature. *International Journal of Industrial Ergonomics*, *76*(January). https://doi.org/10.1016/j.ergon.2020.102928

Brocki, K. C. (2007). *Executive Control Processes: Dimensions, Development and ADHD.* Acta Universitatis Upsaliensis.

Caldwell, J. A. (2005). Fatigue in aviation. *Travel Medicine and Infectious Disease, 3*(2), 85–96. https://doi.org/10.1016/j.tmaid.2004.07.008

Cebola, N., & Kilner, A. (2010). When are you too tired to be safe? Exploring the construction of a Fatigue Index in atm. *Contemporary Ergonomics and Human Factors 2010*, 381–390.

Chang, Y. H., Yang, H. H., & Hsu, W. J. (2019). Effects of work shifts on fatigue levels of air traffic controllers. *Journal of Air Transport Management*, *76*, 1–9. https://doi.org/10.1016/j.jairtraman.2019.01.013

Chen, M. L., Lu, S. Y., & Mao, I. F. (2019). Subjective symptoms and physiological measures of fatigue in air traffic controllers. *International Journal of Industrial Ergonomics*, *70*, 1–8. https://doi.org/10.1016/j.ergon.2018.12.004

Craig, A., & Tran, Y. (2012). The Influence of Fatigue on Brain Activity. In G. Matthews, P. A. Desmond, C. Neubauer & P. A. Hancock (Eds.), *The handbook of operator fatigue* (pp. 185-196). Ashgate.

Dasari, D., Shou, G., & Ding, L. (2017). ICA-Derived EEG correlates to mental fatigue, effort, and workload in a realistically simulated air traffic control task. *Frontiers in Neuroscience, 11*, 297. https://doi.org/10.3389/fnins.2017.00297

DeLuca, J. (Ed.). (2005). *Issues in clinical and cognitive neuropsychology. Fatigue as a window to the brain*. MIT Press.

DePoy, E., & Gitlin, L. N. (2016). Collecting Data Through Measurement in Experimental-Type Research. In *Introduction to Research* (pp. 227–247). Elsevier. https://doi.org/10.1016/b978-0-323-26171-5.00017-3

Dinges, D. F., & Grace, R. (1998). *PERCLOS: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance*. US Department of Transportation, Federal Highway Administration, Publication Number FHWA-MCRT-98-006.

Di Stasi, L. L., McCamy, M. B., Macknik, S. L., Mankin, J. A., Hooft, N., Catena, A., & Martinez-Conde, S. (2014). Saccadic Eye Movement Metrics Reflect Surgical Residents' Fatigue. *Annals of Surgery*, *259*(4). https://doi.org/10.1097/SLA.0000000000000260

Federal Aviation Administration. (2013). *Fatigue Risk Management Systems for Aviation Safety AC for Aviation Safety.* Advisory Circular. No: 120–103A. June 2013.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Field, A. (2016). *Introducing the Linear Model*. Retrieved May 16, 2021, from https://www.discoveringstatistics.com/repository/linearmodels.pdf

Fotiou, F., Fountoulakis, K. N., Goulas, A., Alexopoulos, L., & Palikaras, A. (2000). Automated standardized pupillometry with optical method for purposes of clinical practice and research. *Clinical Physiology*, *20*(5), 336–347. https://doi.org/10.1046/j.1365-2281.2000.00259.x

Garson, G. D. (2013). Fundamentals of hierarchical linear and multilevel modeling. In *Hierarchical linear modeling: Guide and applications, 3-25*.

Golz, M., Sommer, D., Trutschel, U., Sirois, B., & Edward, D. (2010). Evaluation of fatigue monitoring technologies. *Somnologie, 14*(3), 187–199. https://doi.org/10.1007/s11818-010-0482-9

Gunzelmann, G., & Gluck, K. A. (2009). An Integrative Approach to Understanding and Predicting the Consequences of Fatigue on Cognitive Performance. *Cognitive Technology, 14*(1), 14-25.

Hancock P. A., Desmond P. A., & Matthews, G. (2012). Conceptualizing and Defining Fatigue. In G. Matthews, P. A. Desmond, C. Neubauer & P. A. Hancock (Eds.), *The handbook of operator fatigue* (pp. 63-74). Ashgate.

Hancock, P. A., Volante, W. G., & Szalma, J. L. (2016). Defeating the Vigilance Decrement. *IIE Transactions on Occupational Ergonomics and Human Factors, 4*, 151–163. https://doi.org/10.1080/21577323.2016.1178191

Haubert, A., Walsh, M., Boyd, R., Morris, M., Wiedbusch, M., Krusmark, M., & Gunzelmann, G. (2018). Relationship of event-related potentials to the vigilance decrement. *Frontiers in psychology, 9*, 237. https://doi.org/10.3389/fpsyg.2018.00237

Helander, M. (2005). *A guide to human factors and ergonomics*. Crc Press.

Hirshkowitz, M., & Sharafkhaneh, A. (2017). Evaluating Sleepiness. In M. Kryger, T. Roth, W. C. Dement (Eds.), *Principles and Practice of Sleep Medicine* (pp. 1651-1658). Elsevier. https://doi.org/10.1016/b978-0-323-24288-2.00169-0

Holmqvist, K. and Andersson, R. (2017). *Eye tracking: A comprehensive guide to methods, paradigms and measures*. Lund Eye-Tracking Research Institute.

ICAO. (2016). *Doc 4444 air traffic management PANS-ATM*. International Civil Aviation Organization.

John D. Cook. (n.d.). *Linear interpolator*. Retrieved May 16, 2021, from https://www.johndcook.com/interpolator.html

Johnston, M. P. (2014). Secondary Data Analysis: A Method of which the Time Has Come. *Qualitatve and Quantative Methods in Libraryes (QQML)*, *3*, 619–626.

Kaida, K., Takahashi, M., Åkerstedt, T., Nakata, A., Otsuka, Y., Haratani, T., & Fukasawa, K. (2006). Validation of the Karolinska sleepiness scale against performance and EEG variables. *Clinical Neurophysiology*, *117*(7), 1574–1581. https://doi.org/10.1016/j.clinph.2006.03.011

Kilgore, W. D. (2010). Effects of sleep deprivation on cognition. *Progress in brain research, 185*, 105-129. https://doi.org/10.1016/B978-0-444-53702-7.00007-5

Konica Minolta. (n.d.). *CRI Illuminance meter CL-70F*. Retrieved May 16, 2021, from https://www5.konicaminolta.eu/en/measuring-instruments/products/light-display-measurement/illuminance-colour-meters/cl-70f/introduction.html

Kuo, J., Lenné, M. G., Myers, R., Collard-Scruby, A., Jaeger, C., & Birmingham, C. (2017). Real-Time Assessment of Operator State in Air Traffic Controllers Using Ocular Metrics. *Proceedings of the Human Factors and Ergonomics Society 2017 Annual Meeting*, 257–261. https://doi.org/10.1177/1541931213601547

Langan-Fox, J., Sankey, M. J., & Canty, J. M. (2009). Human factors measurement for future air traffic control systems. *Human factors, 51*(5), 595-637.

Lewis, M. (2007). Stepwise versus Hierarchical Regression: Pros and Cons. *Southwest Educational Research Association 2007 Annual Meeting*, 1-30.

Li, F., Chen, C. H., Xu, G., & Khoo, L. P. (2020). Hierarchical Eye-Tracking Data Analytics for Human Fatigue Detection at a Traffic Control Center. *IEEE Transactions on Human-Machine Systems*, *50*(5), 465–474. https://doi.org/10.1109/THMS.2020.3016088

Matthews, G., Desmond P. A., Neubauer, C., & Hancock P. A. (2012). An Overview of Operator Fatigue. In G. Matthews, P. A. Desmond, C. Neubauer & P. A. Hancock (Eds.), *The handbook of operator fatigue* (pp. 3-24). Ashgate.

McIntire, L. K., McIntire, J. P., McKinley, R. A., & Goodyear, C. (2014a). Detection of vigilance performance with pupillometry. *Eye Tracking Research and Applications Symposium (ETRA)*, 167–174. https://doi.org/10.1145/2578153.2578177

McIntire, L. K., McKinley, R. A., Goodyear, C., & McIntire, J. P. (2014b). Detection of vigilance performance using eye blinks. *Applied Ergonomics*, *45*, 354–362. https://doi.org/10.1016/j.apergo.2013.04.020

Meyer, L. (2020). *Final report SafeTower*. LFV System & Utveckling Report D-2020-210735.

Miller, J. C. (2012). An Historical View of Operator Fatigue. In G. Matthews, P. A. Desmond, C. Neubauer & P. A. Hancock (Eds.), *The handbook of operator fatigue* (pp. 25-44). Ashgate.

Miller, L. A., & Lovler, R. L. (2016). *Foundations of psychological testing: A practical approach* (5th ed.). Sage Publications, Inc.

Muscio, B. (1921). Is a fatigue test possible?. *British Journal of Psychology*, *12*(1), 31.

Najjar, R. P., Wolf, L., Taillard, J., Schlangen, L., & Salam, A. (2014). Chronic Artificial Blue-Enriched White Light Is an Effective Countermeasure to Delayed Circadian Phase and Neurobehavioral Decrements. *PLoS ONE*, *9*(7), 102827. https://doi.org/10.1371/journal.pone.0102827

Nealley, M. A., & Gawron, V. J. (2015). The Effect of Fatigue on Air Traffic Controllers. *International Journal of Aviation Psychology*, *25*(1), 14–47. https://doi.org/10.1080/10508414.2015.981488

Noor, N. M., Al Bakri Abdullah, M. M., Yahaya, A. S., & Ramli, N. A. (2015). Comparison of linear interpolation method and mean method to replace the missing values in environmental data set. *Materials Science Forum*, *803*, 278-281.

Noyes, F. R., & Barber-Westin, S. D. (2017). Diagnosis and Treatment of Complex Regional Pain Syndrome. In F. R. Noyes & S. D. Barber-Westin (Eds.), *Knee Disorders: Surgery, Rehabilitation, Clinical Outcomes* (2nd ed., pp. 1122–1160). Elsevier. https://doi.org/10.1016/B978-0-323-32903-3.00040-8

Oken, B. S., Salinsky, M. C., & Elsas, S. (2006). Vigilance, alertness, or sustained attention: physiological basis and measurement. *Clinical neurophysiology*, *117*(9), 1885-1901. https://doi.org/10.1016/j.clinph.2006.01.017

Pfleging, B., Fekety, D. K., Schmidt, A., & Kun, A. L. (2016). A model relating pupil diameter to mental workload and lighting conditions. *Proceedings of the 2016 CHI conference on human factors in computing systems,* 5776-5788. https://doi.org/10.1145/2858036.2858117

Samn, S. W., & Perelli, L. P. (1982). *Estimating aircrew fatigue: a technique with application to airlift operations.* School of Aerospace Medicine Brooks Afb tx.

Schleicher, R., Galley, N., Briest, S., & Galley, L. (2008). Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired?. *Ergonomics*, *51*(7), 982-1010. https://doi.org/10.1080/00140130701817062

Seedhouse, E., Brickhouse, A., Szathmary, K., & Williams, E. D. (2020). Fatigue. In *Human Factors in Air Transport* (pp. 23–47). Springer International Publishing. https://doi.org/10.1007/978-3-030-13848-6_2

Sikander, G., & Anwar, S. (2018). Driver fatigue detection systems: A review. *IEEE Transactions on Intelligent Transportation Systems, 20*(6), 2339-2352.

Soares, C. G., Souza, J. K. S., Ginani, G. E., Pompéia, S., Tierra-Criollo, C. J., & Melges, D. B. (2013). Identification of drowsiness and alertness conditions by means of Spectral F-Test applied to pupillometric signals. *Journal of Physics: Conference Series, 477*(1). https://doi.org/10.1088/1742-6596/477/1/012027

Straussberger, S., & Schaefer, D. (2007). Monotony in Air Traffic Control. *Air Traffic Control Quarterly*, *15*(3), 183–207. https://doi.org/10.2514/atcq.15.3.183

Svensson, Å. (2020). *Human-automation teamwork: Current practices and future directions in air traffic control*. Linköping University Electronic Press.

Swedish Research Council. (2017) *Good research practice*. Stockholm: Swedish Research Council.

Tello, F., Mateos, A., Jiménez-Martín, A., & Suárez, A. (2018). The Air Traffic Controller Work-Shift Scheduling Problem in Spain from a Multiobjective Perspective: A Metaheuristic and Regular Expression-Based Approach. *Mathematical Problems in Engineering, 2018*. https://doi.org/10.1155/2018/4719178

Tobii. (n.d.) *Tobii Pro Glasses 2*. Retrieved April 7, 2021, from https://www.tobiipro.com/siteassets/tobii-pro/brochures/tobii-pro-glasses-2-brochure.pdf/?v=6

Tremaine, R., Dorrian, J., Lack, L., Lovato, N., Ferguson, S., Zhou, X., & Roach, G. (2010). The relationship between subjective and objective sleepiness and performance during a simulated night-shift with a nap countermeasure. *Applied ergonomics, 42*(1), 52–61. https://doi.org/10.1016/j.apergo.2010.04.005

Triyanti, V., Azis, H. A., Iridiastadi, H., & Yassierli. (2020). Workload and Fatigue Assessment on Air Traffic Controller. *IOP Conference Series: Materials Science and Engineering*, *847*(1). https://doi.org/10.1088/1757-899X/847/1/012087

Trutschel, U., Sirois, B., Sommer, D., Golz, M., & Edwards, D. (2011). PERCLOS: An Alertness Measure of the Past. *PROCEEDINGS of the Sixth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 172–179. https://doi.org/10.17077/drivingassessment.1394

van der Linden, D. (2011). The urge to stop: The cognitive and biological nature of acute mental fatigue. In P. L. Ackerman (Ed.), *Decade of Behavior/Science Conference. Cognitive fatigue: Multidisciplinary perspectives on current research and future applications* (pp. 149–164). American Psychological Association.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors*, *50*(3), 433–441. https://doi.org/10.1518/001872008X312152

Watson, A. B., & Yellott, J. I. (2012). A unified formula for light-adapted pupil size. *Journal of vision*, *12*(10), 12-12. https://doi.org/10.1167/12.10.12

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2012*). Engineering psychology and human performance* (4th ed.). Psychology Press.

Wilkinson, V. E., Jackson, M. L., Westlake, J., Stevens, B., Barnes, M., Swann, P., Rajaratnam, S. M. W., & Howard, M. E. (2013). The accuracy of eyelid movement parameters for drowsiness detection. *Journal of Clinical Sleep Medicine*, *9*(12), 1315–1324. https://doi.org/10.5664/jcsm.3278

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in quantitative methods for psychology*, *8*(1), 52-69.

Yamanaka, K., & Kawakami, M. (2009). Convenient Evaluation of Mental Stress With Pupil Diameter. *International Journal of Occupational Safety and Ergonomics*, *15*(4), 447–450. https://doi.org/10.1080/10803548.2009.11076824

Yamada, Y., & Kobayashi, M. (2018). Detecting mental fatigue from eye-tracking data gathered while watching video: Evaluation in younger and older adults. Artificial Intelligence in Medicine, 91, 39–48. https://doi.org/10.1016/j.artmed.2018.06.005

Zhang, J., & Chen, D. (2001). Interpolation calculation made EZ. *14th Annual Conference Proceedings, NorthEast SAS Users Group NESUG, Baltimore, MD.*

Zhang, X., Yuan, L., Zhao, M., & Bai, P. (2019). Effect of fatigue and stress on air traffic control performance. *ICTIS 2019 - 5th International Conference on Transportation Information and Safety*, 977–983. https://doi.org/10.1109/ICTIS.2019.8883823

Zuraida, R., & Iridiastadi, H. (2011). Psychomotor Vigilance Task as a Measure of Performance-Based Fatigue. *Proceedings 7th International Seminar on Industrial Engineering and Management*, 53-57.

Åkerstedt, T., & Gillberg, M. (1990). Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience*, *52*(1-2), 29-37. https://doi.org/10.3109/00207459008994241

TRITA TRITA-CBH-GRU-2021:126