The Future Railway Control Room:

A Practical Framework for Control Room Design Practices

Masters Thesis in Applied Cognitive Psychology

06/08/2021

Author:

Shea Gilmore 6599397

Supervisors:

Niilo Valtakari, Utrecht University

Julia Lo, ProRail B.V.

MSc Applied Cognitive Psychology

Faculty of Social and Behavioural Sciences

Thesis, 27.5 ECTS

Assessors: Niilo Valtakari and Jana Klaus

Utrecht University

Abstract

Socio-technical domains with control rooms, such as the Railway Industry, currently face a gap between the implementation of rigorous user-oriented design and the pace at which digitalization is required. Research has shown that quantitative Human Factors research techniques through investigation of constructs such as Situation Awareness, can be useful, if not essential, to user-oriented control-room design processes. This thesis consists of two parts, the first of which is a conceptual framework based on a literature review, that focuses on the interplay between different aspects of control room design and shows how these aspects could be reconfigured to enable more user-oriented control-room design. The second part of this thesis is a case-study, which aims to investigate the advantages of integrating quantitative, Human Factors techniques into the control-room design process by demonstrating the usefulness of a select range of quantitative, human factors, research techniques. Building on this conceptual framework, the study asks, what is the plausibility of integrating quantitative human factors techniques into usability research in control rooms?

Based on a review of various quantitative situation awareness measuring techniques, a simulation with an existing and a novel interface for train traffic controllers was conducted, and quantitative data was collected to assess various metrics through eye tracking, interval popup queries and a series of self-report questionnaires. Analysis of the ascertained data indicates that eye-tracking measures are representative of situation awareness when looking at associations with self-report and probe query measures. The results demonstrate that quantitative, human factors research techniques are useful in the context of control room design and should be implemented more extensively. Given the results of the case study in the current thesis it is recommended that future research further investigates associations between quantitative research techniques such as eye tracking measures and probe query measures which assess situation awareness dynamically when investigating control room design.

Table of Contents

1 Introduction

1.1 Background

Control rooms are a critical feature in many workplaces involved in safety monitoring and logistical organization and have a vast array of setups and uses depending on the industry. Control rooms are utilized by a variety of industries such as the railway, marine, air, space-traffic-control, security, military, and mass-transportation sectors (Bennet,1993). Control rooms are likewise used in varying manners throughout these varying industries. In nuclear power plants there are generally reactors with turbines, each of which has its own control room, inclusive of sound alarms stationed by the control boards that have several different systems (Andersson, 2010). Petrochemical control rooms have traditionally existed as a series of control rooms with separate functions, each of which has a specialized setup (Kwekkeboom, 2012). In the case of naval operations, maritime ship bridges are the rooms from which ships are commanded and controlled, and often have one to several screens presenting an array of essential information (Hareide & Ostnes, 2017). In the space sector, satellite control room operators must operate physical parts of the satellite with a software that utilizes three individual screens, and have larger, more holistic screens present at the center of the room (Berhaupt et al., 2007). And finally in the railway industry controllers have a several screen setup with monitors positioned in a curved manner to allow oversight into all areas, with larger and higher resolution monitors in more modern control rooms, and a phone, and sometimes a radio microphone, for communication with the drivers (Lischke et al., 2018). The systems present on these computers utilize self-control mechanisms, and allow controllers to register events, have remote control of the trains and assist drivers in conducting their operations in a safe manner (Kornaszewski, 2018). Generally, control room design has indicated a proclivity towards digitization as technology has developed. Because controllers largely operate systems remotely, the interface design is critical to their job because it serves the purpose of conveying their highly important and safety-oriented tasks (Anderson, 2010).

Specifically in the Netherlands, traffic controllers (TTCs) working in railwayy control rooms work with an eight-screen system which allots different applications by function on the different screens, with the most active ones being centralized (Lo et al., 2017). In the railway industry TTCs are required to monitor operations and identify abnormalities, alongside ensuring that railway proceedings such as track logistics, track maintenance, and infrastructure and transport events are apprehended and attended to in a timely manner, and thus play an integral role in the railway operation industry. Given the importance of the role of TTC's, it is crucial that TTC's can work in an appealing environment that allows them to execute tasks efficiently, effectively and with minimal error and maximal reliability.

In the Netherlands the railway sector is expected to experience a sharp incline in train traffic soon (Van Leijen, 2018), meaning that Dutch railway traffic control rooms will need to be adapted and updated to cope with this change, with future traffic control room design needing both coherency and precision. Effective railway control room design should ensure that salient features are presented well through effective system design (Crawford, Toft and Lift, 2013). A range of prototypes and setups have previously been investigated to promote the progression of railway development in the Netherlands, to explore what the future workplace of a TTC could look like

(Lo et al., 2017). Much of this prototyping has employed usability techniques to assess prototype efficacy and has been conducted by the infrastructure company Prorail. At present however, there is still a want for more widespread and unified implementation of human factors methods in control room environments, particularly in the railway industry (Crawford, Toft and Kift, 2013). Also, while human factors construct such as mental workload and situation awareness have been extensively researched in a variety of domains, they have not had a uniform uptake in safety critical domains such as railways control rooms, although they are deemed to have vast potential for modelling human navigation of a complex environment (Kovedsi et al., 2018). It has thus been proposed that traffic-control interface-design must be more extensively researched to support automatic cognition and decision making in dynamic environments (Kauppi et al, 2006). Moreover, an emphasis on human-in-the-loop (hereafter HITL) and user centered design in railway control rooms through rigorous scientific investigation has been emphasized. These are thought to be crucial factors for effective operator decision making under pressure (Kauppi et al., 2006). In the same line, increased digitization of safety operator environments has been thought to result in lowered operator situational awareness, and uneven cognitive workload distribution because environments have been found to fluctuate between varying degrees of repetitive, automated tasks, and high-pressure situations (Hareide & Ostnes,2017), making more in-depth investigation into the interplay between these human factors constructs and design a necessity.

1.2 Purpose and Aim

When accounting for this dynamic state of control room operations at present, and the process of rapid modernization as well as the want for optimization of control rooms, further investigation of control room design, through a usability lens, is necessary.

The current thesis will thus present:

- **A conceptual framework** that presents an overview of system design and investigates potential new methods for control room design analysis, focusing specifically on the role that human factors can play in this. Subsequently this conceptual framework will be utilized in a case study that addresses human factors measures relevant to assessing aspects of control room design and usability.
- **A case study** to illustrate the conceptual framework, as a steppingstone for future control room systems research and design, as well as to address the future state of the Dutch railway industry, which will be the context in which the case study is conducted. The study will investigate a new interface for TTCs and the efficacy of quantitative tooling.

The overall intention of the current thesis is to contribute to future control room design by exploring quantitative research approaches in a control room systems context, and to investigate the theory they are based upon.

It is hypothesized that quantitative research methods can be used to assess control room systems design more extensively. The results should be useful for human factors psychologists working in the ergonomics field. Given that ProRail has previously tested a series of prototypes, the current research will build on this research by assessing to what extent quantitative usability methods and tooling are feasible in a control room environment, and representative of human factors

constructs. Thus, it will ask, what is the plausibility of integrating quantitative human factors research techniques more extensively into design methods in control rooms?

The current research will likewise add more dimensionality to control room research approaches, by asking ‘what are the current, ubiquitous methods for analyzing in a control room design?’. This appraisal will be integrated into a conceptual framework and will in turn assist in paving the way for incorporating new approaches for control room usability research.

This first chapter of the thesis was intended to introduce the control room environment and give an overview of design shortcomings and potential solutions, as well as a broad description of what to expect from this thesis. The upcoming second chapter of this thesis will center around a literature review which will assess the current state of Human Factors as a field, and the ways in which it is highly applicable to control room design. The two parts of this chapter, 2.1 and 2.2 focus on relevant human factors constructs and aspects of system design pertaining to control room environments. Building on this the latter section of the second chapter, 2.3, realizes these many elements as a conceptual framework for future control room design processes. The subsequent and final third chapter of this thesis consists of a case study conducted in an applied research environment with train traffic controllers. The study uses the conceptual framework as a basis and consists of a series of quantitative and qualitative measures to explore the nature of control room interface design elements and processes.

2. Human Factors and Safety Critical Domains

Human factors (HF) and ergonomics is the multi and interdisciplinary domain whereby scientific and design theory and methods are applied to optimize human and object interactions and outcomes in a range of settings (Karwowski, 2012). Historically the driving force behind the accelerated development of the field of human factors was the development of technology. In contrast to this, it is thought that proactive, verified design methods with the intent of optimization should drive the development of technology in the future (Karwowski, 2012). For this to occur, a larger emphasis should be placed on fitting design to work with natural limitations such as human perception and operator fatigue so that it operates in a bottom-up manner, rather than designing without consideration of natural constraints, and disregarding technological ecology (which in the case of control rooms would refer to humans and the technology/ machines at hand) and designing in a top-down manner (Karwowski, 2012).

The Human Factors and Ergonomics Society has advocated for rigorous application of knowledge with regards to human characteristics, but the definition has often played out in a limited scope and primarily dealt with spatial domains, such as ergonomic alteration of workspaces to prevent poor posture for example (Karwowski, 2012). Karwowski (1997) introduced the notion of human compatible systems investigation and design, whereby objective and direct (e.g., psychophysiological measurements) next to subjective and indirect (that is qualitative) metrics are applied to establish more clearly which specific human outcomes are the result of specific HF contexts. Moreover, it is important to consider that the (re)design of control rooms, is contextually dependent upon the symbiosis of the human, the technology, and the organizations in which these changes are being implemented in which case a more holistic approach may not also be feasible when considering time and monetary constraints, or safety procedures and organizational policies (Karwowski, 2012; Johnsen & Veen, 2007). Thus, in line with this notion, the current thesis will review the state of human factors implementation and investigation in complex sociotechnical domains where safety critical work is performed. Furthermore, this chapter is intended to illustrate the many components that are relevant to improved control room design investigation. Each of these components has been subjected to literature review and are outlined as single points below. At the end of this chapter the interplay between these components will be illustrated with a conceptual framework diagram.

2.1 Systems Design

Historically, system design in a human factors context has involved a reductionist approach that has been operationalized without extensive considerations for behavioral aspects of the systems at hand and has instead primarily supported the application of engineering and sociotechnical systems (Czaja & Nair, 2012). The systems approach to system design involves the consideration of all system components and their interplay with regards to the system goals for evaluating a specific concept (Czaja & Nair, 2012). However, systems currently in place in safety critical domains are usually too complex for the integration of a set of individual subsystems to support user experience and have resulted in suboptimal performance and

design (Sage & Rouse, 2009). Likewise, systems have been found to fall short of the mark in novel, unprecedented situations involving people (Gorman et al., 2010). These findings highlight both the dire need for consideration of the human element to cope with these shortcomings and the need for coordinated and critical implementation of subsystems.

There have been general trends towards control room modernization and redesign across different sectors, to ensure that processes become more effective, efficient and standardized (Kwekkeboom, 2012). As previously stated, the integration of human factors into control room design has been proposed to be an important, if not essential, step in future control room interface design (Dos Santos et al., 2007). In the petrochemical industry, there has been a push to centralize control rooms into one facility, to increase efficiency and improve communications (Millner et al., 2016), as well as a push for the implementation of a unified safety system (Kwekkeboom, 2012). Nuclear power plant control rooms have seen increased digitalization of traditionally analogue systems throughout the process of modernization (Salo & Savioja, 2006), as well as a push for larger screens usage to attain a more holistic perspective of processes occurring (Harrefors, 2008). In railway sectors there has been emphasis placed on user centered design, with clean and integrated presentation of information being a key target, as well as homing in on interfaces optimization for controller use (Elliot, 2010). Likewise challenges have been presented with the increasing digitization of the environment with a want for system redesign arising (Andersson et al., 1999). Control room operators thus often work with complex systems in a fluctuating environment, making their profession a highly skilled and specialized one (Kluge et al., 2014). Like any specialized profession, this can lend to difficulties in recruiting and retaining operators.

Specifically in the Railway sector, insufficient integration of human factors into safety management has been found to be a significant contributory factor for human to system incompatibility (Crawford et al., 2013). Furthermore, in many countries there is no standardized human factors tooling established for use in the context of a Railway control room setting (Crawford et al., 2013). This has been thought to occur because of a lack of system unification between different types of operations (such as metropolitan versus regional) as well as top-down design that hasn’t sufficiently considered MMI clashes and differing organizational standards between and within countries whereby there is no distinct human factors integration policy (Crawford et al., 2013). Boring and Joe (2015) propose that control room modernization guidelines are missing the mark by focusing on regulatory processes alone, rather than assimilating ergonomics and usability baseline evaluations into their framework. These two evaluation types consider systems currently in use and assess the operator performance and potential design shortcomings respectively, to better inform design of replacement, subsequent systems (Boring & Joe, 2015).
Furthermore, lack of ergonomics contribution in early control room design phases has too often resulted in installation and instrumentation by technicians that have not considered the requirements and procedures of the operator's job (Pikkar, 1992). It has also been emphasized that future control room system design needs to be human-centric and based on clarified workflow and user experience to allow tools and functions to coevolve (Mentler et al., 2018). In the same line, Kovedsi et al (2017) place particular emphasis on the efficacy of eye-tracking

measures for measurement and evaluation of control room modernization, suggesting that eye-tracking metrics are an integral addition to the diverse set of tools that could be used to facilitate future human-system interplay and system design in control rooms.

2.2.1 Automation

2.2.1.1 Automation and System Design

Varying degrees of automation are widespread in sociotechnical setups at present (Mentler et al., 2018) and have been implemented in safety critical fields to reduce the overall mental workload of the operator to take care of mundane, repetitive tasks (Bannon, 1991) and to increase efficiency and system productivity (Andersson, 2010). Ideal automation in a human machine interface functions symbiotically and uses features as a memory aid to guide operators working under time pressure (Boy, 1995). Moreover, because MMI has been implemented at an accelerated pace into control rooms, and because there is much more flexibility in design than there is for traditional hardware setups, many automated systems lack a standardized design framework, which would be ideal for implementation (Kitamura et al., 2005). Measuring these setups is a very complex task with an unprecedented number of display possibilities and volume of communication, as well as an increased number of formats and control modalities (such as verbal commands) (Czaja & Nair, 2012). Also, latency in response to emergencies due to poor alert signaling and thus poor perception of it from operators and difficulty in problem diagnosis, combined withs the fact that these activities take time themselves before the emergency can be attended to is where human error comes in to play because of the out of the loop performance problem (Endsley & Rodgers, 1996). Likewise, usability comes into play in high responsibility MMI environments given that intuitive representation of complex systems displayed information is remarkably difficult (Endsley & Rodgers, 1996).

2.2.1.2. Automation and human error

Historically in literature there has been a tendency to attribute the error in MMI environments more towards human error, but it is more important to look more broadly to the overall mechanisms of the environment at hand (Young & Stanton, 2002). It has been found that poor automation system response arising out of outdated control equipment results in ad-hoc adjustments by human operators, which should not be viewed as violations of reliability but rather conditioned reactions to cope with unanticipated events in the system design (DeCarvalho, 2005). It has likewise been found in traffic operator environments that all too often system developers automate easy, monotonous tasks, and leave traffic operators with a mixed bag of cognitively heavy tasks that after a period of sustained attention will promote fatigue (Dadashi et al., 2013; Endsley & Rodgers, 1996). Mental workload in the automation environment has been shown to be altered, with work underload due to automation can be as detrimental and mental work overload, with environments that fluctuate rapidly between the two being especially extreme (Young & Stanton, 2002). Moreover, because automated systems at present can have more rules, decision pathways, and multiple controller regions it can be said that operators must grapple with this burdening their MW level when interacting with a new system, because lack of standardization requires operators to learn each design vision anew (Czaja & Nair, 2012). With regards to situation awareness, it is proposed that three mechanisms are at work with the impact automation has upon it, these being namely that

automation alters the way in which operators receive feedback and each system varies in the way it is presented (thus operators must be aware of this), also fluctuations between vigilance and complacency occur more because operators must be continuously monitoring between spurts of activity, and finally automation means that operators must take a more passive role than prior to automation implementation leading to a more complacent, lowered state of situational awareness overall (Endsley, 1996). It is proposed that especially in the case of railway control rooms, intricate MMI workplace structures and processes need to be accounted for in the planning of technological implementation (Dardashi et al., 2013).

Given that control rooms are likely to contain even higher levels of integrated automation, because operators will need to cover more expansive track areas, automation is an important element to consider when assessing system design of future control rooms. The concept of levels of automation in a control room are important to the system design of the control room, given that there is a risk that the tasks left for the human to control are either too simple or too complex (as previously mentioned), meaning that often tasks which are very simple are too expensive and seldom occurring to automate, and tasks which are very complex often need updating and are too difficult to program (Andersson, 2010). If the human element level of participation in the overall system is too little most of the time (that is, when performing more passive and repetitive tasks) but alertness is suddenly needed when an emergency occurs, operators will have to suddenly fluctuate between complacency and vigilance (Andersson, 2010). Likewise, a sense of control has been found to be intrinsic in control room operator job satisfaction with too low a level leading to disenfranchisement to occupation (Reiman & Oedewald, 2009). Also trust in automation is important in avoiding misuse and disuse (Lee & See, 2004). Specifically in the context of train traffic management technology, human automation design and continuous integration efforts have been highlighted as mechanisms of risk to both safety and railway operations functionality (Crawford & Kift, 2018). Several key trends have been found to be relevant to these automation issues in the railway control rooms, which are namely, large scale changes in safety procedures because of technological change, end user uptake of technological systems with rapidly increasing complexity, increasingly sedentary work and increasing demand for data analytics to support process optimization (Crawford & Kift, 2018). Moreover, investigation has shown that end user occupational requirements are not being acknowledged or met due to a lack of time for implementation and training (Crawford et al., 2014). Likewise, organizational factors such as slow adoption of technology, lack of accountability and top-down project implementation were also shown to diminish the adoption of technology to support resilient organizational processes in the railway sector (Crawford et al., 2014). It is therefore important to examine the interplay these elements of automation have when looking at control system learnability and design (Bannon, 1991; Andersson, 2010).

2.2.2 Prototyping

An effective method for ensuring that a future system design is effective is prototyping, which is a widespread method used in the display and testing of a new interface or system. Prototyping typically involves the testing of concepts such as new functionalities and interface, in order to develop and pilot new designs before they are implemented (Barbieri et al., 2013; Walker et al.,

2014). Prototyping can utilize a selection of physical archetypes to allow participants to compare features of different options (Barbieri et al., 2013). On the other hand, virtual prototype testing is a method which allows for decentralized system design and development and allows participants to manipulate the model through a screen (Kuutii et al., 2001). Prototypes can differ in terms of high and low fidelity between and within their categories (Rudd et al., 1996), however it has been found that level of fidelity is not necessarily integral to uncovering usability issues in the process of system design (Uebelbacher et al., 2014).

2.2.2.1. Low fidelity prototyping

Low fidelity prototyping is advantageous because designers can easily and affordably generate prototypes that can be discarded by clients without a big loss of resources and effort, whilst also allowing designers to gauge user's responses with a raw version of a prototype before investing energy into creating a high fidelity version (Lewis, 2012). Low fidelity prototypes are created less for appraising interaction and more for depiction of alternatives, setups, and conceptualizations to inform users and clients of the overall basis for addressing the problem statement with this rough design (Camburn et al., 2017). Low fidelity prototyping is pertinent to agile design methodologies because they allow designers to draw from a large pool of iterations that evolve and adapt over time to eventually form a higher fidelity prototype (Camburn et al., 2017). When prototyping is performed well, the conceptualisation of a high-fidelity prototype features will be grounded in previous iterations of low fidelity prototypes (Camburn et al., 2017). Low fidelity prototyping done with simple tools such as pencil and paper has been theorised to be highly effective because designers don't get bogged down in the technicalities of navigating software and can thus have a higher output of iterations, making the path to the higher fidelity prototype more efficient (Camburn et al., 2017).

2.2.2.2 Mockups and High-Fidelity Prototyping

Wireframes, storyboards, wire flows and mockups are important terms in the prototyping lexicon. Wireframe refers to either the initial (low fidelity, usually static) series of sketches representing the interface or environment setup being designed serving the purpose of a blueprint, or high-fidelity wireframes which are used to test aspects or interface features usually on a screen or between multiple screens (Camburn et al., 2017). Storyboards utilize simple software like PowerPoint (but can also use specialized software or even be done in analog format) and are a sequential flow of wireframes to establish a narrative for the user for the task at hand and assist in elucidating features in the early stages of development (Camburn et al., 2017). Wireflows also represent the sequential flow of wireframes but rather than a linear representation it identifies user flow multidimensionally, allowing for multiple navigation paths and representation of complex interactions and user paths (Camburn et al., 2017). Mockups are a high fidelity, nonfunctional, static representation of a design that portray the aesthetic of the final prototype and are often utilized after the wireflow stage of prototyping (Sauer et al., 2010). These prototyping methods are pertinent to interactive design because they allow for deliberated evolution of a system or product, ensuring that the most pragmatic outcome is reached (Camburn et al., 2017), making higher fidelity testing more desirable in technical environments where avoidance of human error is key to ensuring systems work optimally (Otto & Wood, 2001). Simulation environments are applicable when other elements come into play,

such as hardware, and a higher fidelity, emulated approach is required (Camburn et al., 2017). Simulations are an important resource for informing system design in sociotechnical systems, where they can be used to test out the knowledge and skills of operators, and use the results to inform change, as well as allowing research teams to engage with stakeholders in company contexts to ensure open communication about system design decisions based on the users (Lo, 2020).

2.2.2.3 Human in The Loop Mockup

In safety critical environments a common method for assessing usability involves testing a prototype in a human in the loop (hereafter HITL) mockup, which is a methodology that was established to both compensate for the constraints of human performance, as well as to enhance human performance outcomes (LeBlanc, 2014). HITL prototyping is utilized to both verify performance and safety standards as well as for testing different iterations of a prototype at varying levels of fidelity before actual implementation (LeBlanc, 2014). HITL simulations require validated simulator tools to create an accurate model which represents interactions between elements which would occur in the actual work environment. Likewise, operators with specific knowledge of local features should be engaged in the research to strengthen HITL simulation validity (Lo, 2020).

It has likewise been theorized that in safety critical environments quantitative data is required to accurately model operator performance and develop a system that accounts for HITL, so that the system can effectively interpret and adapt to operator performance and is thus inherent to human factors techniques (Hu et al., 2019). The reason that HITL prototyping is so prevalent in safety critical settings, is that humans are adaptable to the dynamic nature of safety critical situations, and that the adaptable behavior of humans is currently not automatable (Hu et al., 2019). HITL simulators are used to show how higher fidelity, later prototypes with functioning interfaces work correctly and cohesively (Hu et al., 2019).

2.2.2.4. Automated usability prototype testing

The idea of a simulator or prototype with built in, standardized usability assessment tools has long been contemplated as an attractive alternative to traditional usability techniques, since it would be more time efficient and less consumptive of resources (Chang & Dillon, 1997). Traditionally automated usability has been conceptualized as a software which is installed on the computer that automatically runs in the operating system in the background and can capture the user's interactions with an array of applications with full dimensionality (Chang & Dillon, 1997). More recently a need for remote automated usability practices has arisen alongside the rapid process of agile software development, with software being developed that both allows prototype creation and usability evaluation simultaneously, alongside remote testing of potential users (Hosseini-Khayat, Hellmann & Maurer, 2010). In mobile applications, automated usability tools have been developed into a recording framework whereby a specific pattern of interactions is recognized and sent to a central server for analysis (Kluth, Krempels & Samsel, 2014). This type of pattern-based GUI testing is usually based on a series of Human-Computer Interaction test patterns that are configurable to different applications (Kluth, Krempels & Samsel, 2014; Dias, 2017). However, not all tools labelled automated usability testing tools are testing for

usability, with many tools depicting one dimensional results such as site abandonment or screens visited without embedding the results into a more specific context (Dicks, 2002). In sociotechnical domains where systems are designed to be technocentric, and basic usability concepts and human centric design have not been adequately considered in the design process (Meier & Merten, 2017), reaching automated usability is a two-step process whereby human centered design and more traditional usability techniques would be the first step to gain insight into the kinds of parameters that might be investigated. Likewise, it is also important to note that the kind of automated usability that has been implemented in other domains, such as web development, could manifest differently in the sociotechnical domain, and that automated usability practices should be relevant and specific to the domain at hand (for example, train traffic control).

2.2.2.5. Prototyping the future workplace of a train traffic controller

ProRail B.v. is the company where the majority of TTCs in the Netherlands work and is tasked with handling the network railway infrastructure activities, railway traffic flow and rail allocation, and is thus the employer of the vast majority of TTC's in the Netherlands. ProRail has previously investigated a range of prototypes to promote railway innovation in the Netherlands. The future workplace of a TTC has been explored by developing a concept of a machine man interface (MMI) that was based on the current system functionality and the current tasks of a TTC, with the workplace of TTC's in Zwolle being used as a pilot (Lo et al., 2017). A diverse range of materials, methods and techniques were used to determine how the ideal workplace could look, and it was found that intuitive interaction with the task environment alongside precise and clean delivery of information and ergonomic aspects of physical space were all integral factors to be addressed (Lo et al., 2017). It was emphasized that future development should be centered around user friendly design, with this phase focusing primarily on envisioning the future workplace and exploring ergonomic elements (Lo et al., 2017). Following these two rounds of a total of six varying hardware setups were assessed (with applications and hardware making the demo up as a whole), and an additional touchscreen which is not present in the current TTC workplace. This phase utilized task analysis, interviews and eye-tracking to determine what the use of the current system is and explored guidelines that had to be fulfilled by the future workplace and resulted in three sketch mockups which bigger screens within the ergonomic guidelines were ideal (Lo et al., 2017).

The next stage of testing the future workplace will be conducted by means of the case study presented in the subsequent chapter of the current thesis, which will center around the comparison of the current interface with a new interface, which is under development for a new system being developed. The case study will examine the efficacy of different human factors methods in control room testing by gaining insight from quantitative data collection (such as eye-tracking, situation awareness pop-up probes and simulation logging), as well as more contextual quantitative methods (such as Likert scales, to assess usability, validity and user experience), and qualitative open-end questions and debriefings.

2.2 Constructs

Commonly utilized constructs in the human factors field are conceptualized as operating on a cognitive level and manifesting themselves through behavioral and psychophysiological outcomes (Durso & Alexander, 2010). They have likewise been shown to be distinct from other kinds of ubiquitous psychological constructs such as memory through both a large body of empirical evidence and theoretical frameworks (Durso & Alexander, 2010; Vidulich, 2012). Like many scientific disciplines, the constructs and their interrelationships form a basis for human factors as a discipline (Vidulich, 2012). Key constructs like mental workload (MW) and situational awareness (SA) evolved out of a need to model the mental processes of operators during man machine interaction (MMI), when the attainment of specific information surrounding operator cognition was not possible (Vidulich, 2012). It has been argued there are many more cognitive constructs can provide insight into complex MMI relations, such as human error metrics, performance measures and decision-making models to name a few (Parasuraman et al., 2008). However, it also has been argued that MWL and SA (they) serve a more operational purpose as tools utilized in predicting outcomes in safety critical environments (DeWinter, 2014). Therefore, the current thesis will thus focus on these two key concepts that are predominant at present in human factors literature.

2.2.1 Task Analysis

Ergonomic task analysis involves stripping a task back to its basic components, that is, investigating who does what, and why, and how (Hollnagel, 2012). Ergonomic task analysis is necessary to assess complex and collaborative activities at their primary components, as well as the manners in which these primary components have interplay and to what extent each component is represented by an autonomous mechanism (Hollnagel, 2012). Also, it is necessary to assess interdependency among human elements as well as interaction in the mastering of complex technological elements (Hollnagel, 2012). In control room environments hierarchical task analysis (HTA) is a widespread task analysis tool that decomposes a primary task into a series of subtasks until an elementary level of tasks is reached, with each subtask having a goal, and each pathway through the HTA portraying a sequence that could occur as the result of an interaction between the TTC/ operator and the system (Hollnagel, 2012). Cognitive task analysis (CTA) is another type of task analysis which assesses the required skillset and the mental demand of a task, with applied cognitive task analysis veering more in the direction of the design world (Millitello & Hutton, 1998), and more general CTAs being focused on high level operations (Schachak et al., 2009). HTA and CTA have been differentiated by their outcomes, with HTA being more goal oriented, and descriptive of system characteristics, and CTA being more attuned to identifying system constraints and potential outcomes (Salmon et al., 2009).

2.2.2 Usability Testing, Interactive Design and User Experience

Usability is a widespread construct that is commonly used in the quest to find optimal design strategies. It is determined by the synergy between the user and tools, in the context of the task and environment (Schackel et al., 2009). Usability assessment is conducted with the aim of conforming visual and tactile aspects of a systems design to make them intuitive to functionality and purpose (Sulaiiman et al., 2009), thus making it a valuable tool for analyzing control room design in safety critical contexts such as the TTCs work environment. Usability research

techniques are employed to probe design efficacy by assessing to what extent users can reach end goals (Schackel et al., 2009). Usability was created because of two diverging necessities, the first of which emerged from the intention to correct deficiencies in existing technology, and the other of which emerged from the intention to intervene early in the design process and ensure design was intuitive to future user's requirements (Lewis, 2012). With regards to human factors and usability specifically, successful testing methods have mostly involved case studies done in an iterative manner, with careful documentation and interpretation of results (Lewis, 2012). Usability is a process integrated into design practices in both Human Factors and User Experience fields to ensure that products are relevant and intuitive to users (Furniss et al., 2018).

2.1.5.1. User Experience

User experience (UX hereafter) is a relevant but distinct field to human factors, that is sometimes confounded with it, and that also utilizes usability as a testing construct and aims to uncover valid pitfalls in design, amend them and test them in an iterative manner (Barnum, 2019). UX was born out of a need for a pragmatic, user centered approach in real world business settings, such as web development (Rajanan et al., 2017). Shortcomings in the UX field have been outlined as a lack of valid research approach, real user participation and/ or a shortfall in moderator skill (Barnhem, 2019). Likewise, a focus on only consistency of outcome rather than effective practices, as well as UX democratization (whereby processes are standardized to a point where methods can no longer be tailored to specific contexts, and quick outcomes are expected) has led seasoned UX practitioners to postulate that UX practice is becoming more diluted and potential downfalls of design could be overlooked (Barnem, 2019). On the flipside there has also been an emphasis placed on the need to standardize UX design to promote ethical practices and consult academic research from fields dealing in cognition and psychology, to avoid UX design that is able to manipulate users in an unsavory manner, such as social media platforms with features like endless scrolling that are designed to be addictive (Gray et al., 2018).

2.1.5.2. User Experience and Human Factors

While both UX and human factors fields focus on adapting environments to be more human centric, UX diverges from Human Factors in that UX practitioners work in contexts primarily consisting of interactive systems which are used in a broad range of contexts such as mobile apps, websites, and physical spaces, whereas human factors practitioners generally work in more technical settings such as safety critical ones (for example, in control rooms) (Furniss et al., 2018). Both UX and human factors fields have been confronted with the issue of being interpreted as solely aesthetic design in some industries, but UX has in general had a higher uptake, likely because it is seen as more pragmatic and implementable, and Human Factors is generally interpreted as being more solely academic and applied in academic and technical settings specifically (Fraser & Plewes, 2015; Hassenzahl, 2008). However, it is important to note that these two fields could be a subset of each other, and that academia has generally conflated the two terms such that they are sometimes used interchangeably and sometimes semantically different because of the different contexts they appear in (Fraser & Plewes, 2015; Hassenzahl, 2008).

Moreover, investigative trends in UX tend towards more qualitative approaches and the human factors field has generally used a combination of both. UX commonly utilizes subjective interpretations of a system or interface which allows for precise and personalized analysis. Current and retrospective 'think aloud' and open-ended queries allow users to clarify their comprehension of, and interaction with an environment or prototype, with task based versus open exploration qualitative setups allowing for different usability perspectives to emerge in real time (Sulaiman et al., 2008; Fabo & Durikovic, 2012; Kuutii et al., 2001). While these methods give a large range of responses that can be difficult to interpret in a concise, resource efficient manner, they allow for a diversified range of responses and exploratory learning and can be useful in fine tuning bottlenecks because users can elaborate on their subjective experiences critically (Sulaiman et al., 2008). Qualitative research methods have been shown to produce problems such as self-report bias and secondary task distraction (Schiessl et al., 2003).

2.1.5.3. Interaction Design

Interaction design is also a widespread term relevant to usability, and is the application of usability techniques, with emphasis placed upon the concept of active users (that is users as agents for whom the design is instrumental in achieving an outcome) and has been employed as a term in the fields of human factors and human computer interaction (Bannon, 1991). Interaction design occurs in both technical and other environments looking at user experience and can be thought of as the confluence of science, engineering, and artistic design practices to reach pragmatic, scientifically sound and effective outcomes (Zimmerman et al., 2007). Affordances are an important example of a notion present in the interaction design field that depict how interaction design is pertinent to technical environments such as control rooms by referring to what a specific design characteristic offers a user (Hartson, 2003). Hornbaeck and Stager (2006) pointed out that the fusion between usability engineering and interaction design has been frequently suboptimal, and suggest that either clear usability evaluation and interaction design roles need to be established to better ensure that no steps in the process of ensuring that relevant concerns fall through the cracks, or that the two types of roles are integrated together and performed by practitioners with knowledge of the significance of both roles in good design.

2.1.5.4. Usability research in the Sociotechnical environment

Sociotechnical environment-based usability research has consisted of both qualitative and quantitative research methods used both separately and in conjunction with one another in control rooms (Greenberg & Buxton, 2008; Sulaiman et al., 2008). Both methods are crucial to safety critical environments because the complexity of the systems and their tasks leaves investigation prone to reliability issues when just one type of method is used (Lewis, 2012). While control rooms previously had usability issues regarding physical distance between equipment and heavy physical strain sometimes, depending on the line of work, currently there is thought to be more issues surrounding information presentation, with such an influx of information being available to operators that a funnel effect is thought to occur whereby only a small amount of information being presented can be processed by the operators themselves (Rassmussen & Laumann, 2014). Interaction design in control rooms at present involves not

just consideration of the user but also the safety procedures (Boring et al., 2005). Thus, it is important that control room usability investigation today promotes effective tactics for presenting salient information and address pitfalls such as lack of unification for aspects such as procedural protocol, color coding and semantic control room elements (Rassmussen & Laumann, 2014).

2.2.3 Situational awareness (SA)

SA is formally understood as the mental blueprint of the operator at a specific point in time in a given work environment (Endsley, 2012). SA is conceptualized as a three-level procedure whereby aspects of an environment are perceived, comprehended for their meaning, and subsequently projected for their future status, and responded to accordingly (Endsley, 1996). Attention is closely intertwined with SA and is a multifaceted construct that is in continual interplay with many cognitive processes, with different types of attention being confined to different parts of the brain in terms of cortical activation (Vidulich & Tsang, 2012). In line with this, safety critical settings such as those experienced by TTCs have environments with subtasks that demand similar attention, creating competition for attentional resources. Also, attention allocation is thought to be somewhat strategic and voluntary, with prioritisation of attention changing with changing task demands in safety critical settings (Vidulich & Tsang, 2012).

Also, it is important to note that SA is closely tied to working memory, and timely perception of patterns and objects are not possible without ease of contact to prior knowledge, and thus both declarative and procedural knowledge retrieval from long term memory must occur by means of working memory (Vidulich & Tsang, 2012). Because of this SA can differ between operators who have accumulated differing levels of knowledge and expertise (which is largely learned, by means of deliberation) over time (Lo, Sehic, Brookhuis & Meijer, 2016; Vidulich & Tsang, 2012). On the same note, experts tend to have more structured knowledge in their long-term memory (through practice and association), so it is more efficient for them to retrieve it (Vidulich & Tsang, 2012). Strategic management is a quality thought to be present in operators who exhibit a high level of SA, and involves continuous chunking, strategizing, and reshuffling of information, alongside the inhibition of non-salient information (Vidulich & Tsang, 2012). Thus, SA itself is a distinct construct that, while related to other more broadly applied psychological constructs, cannot be solely defined by them in a reductionist manner. Moreover, good operator situational awareness can be thought of as the operators and the environment's abilities to facilitate operator sustainment of attention as well as support the operator in coping with and perceiving fluctuating elements in a complex environment.
High responsibility domains, such as those used in the Railway sector by TTCs, require highly complex systems which accumulate a mass of data to execute tasks such as railway traffic management. More specifically, a TTC must first perceive elements of their environment, such as what is occurring on the planning screen and the service screen, where different trains are positioned, what subtasks are waiting to be performed. Level two SA would involve the TTC going beyond being solely aware of these elements, and considering how they are instrumental to specific goals, and comprehending salient features to form a holistic notion of the environment. For example, the TTC may have to perceive what level of danger a specific track

disturbance presents and what they can do to mitigate the seriousness of the situation's outcome. Level three SA would involve the TTC being able to project what will occur in the near future given what disturbance has occurred at present, which is achieved by means of mastering the first two levels of situational awareness and having a knowledge of the situation and dynamics at hand. Good railway TTC system design recognizes that it is beneficial to enhance situational awareness for train traffic controllers, both through physical (for example, by use of haptic feedback) and cognitive facilitation. The integration of highly complex systems usually results in an overload of information conveyance through channels, which is where human error usually comes into play (Endsley, 2012). In the case of high responsibility domains, it is important to recognize that true SA exists within the operators own mental model and design should be centered around this (Endsley, 2012). Thus there is a need necessitated for the investigation into how SA can be better promoted by control room design, and how it can be better used as a tool to inform control room design, given that it is mental workload and situation awareness are considered two of the strongest tools in identifying human-system performance mechanisms (Vidulich & Tsang, 2015) .

2.2.4. Mental workload (MWL)

Mental workload is conceptualized in a variety of ways. The simplest definition conceptualizes it as the mental cost of completing a given task at hand (Fallahi et al., 2016). More elaborately put, it is the balance between the cognitive resources, such as working memory and attention, required by a specific task, and the availability of those resources from the operator completing the task, in a supply and demand type of interplay (Parasuraman et al., 2008). Another conceptualization of MWL is made through two key determinants; the exogenous task demands (such as specified task features like contextual factors, priority and difficulty) and the endogenous supply of cognitive resources (such as planning, memorizing and making and executing decisions), which is mediated by the individual's innate aptitude as well as expertise and skills (Vidulich & Tsang, 2012). Moreover, MWL is thought to be an intervening variable that is inferred from performance measures and is thought to causally influence performance outcomes (Kantowitz, 2000). When MWL is either too high or too low, it is thought to result in error in performance (Kantowitz, 2000; Foy & Chapman, 2018).

It is important to recognize that MWL and SA have a high degree of interplay in complex, safety critical settings, and there are many elements to cognition simultaneously interacting with one another (Vidulich & Tsang, 2012). While SA refers to the state of vigilance that the operator can sustain in the control room and the cognitive mechanisms at work, MWL generally refers to the workload demand relative to the resources available to cope with it and the quantitative amount of these resources required (Stephens et al., 2015; Wickens, 2003). Moreover, SA can be thought of as competing with MWL in the case that performance may be sacrificed to maintain longevity of SA, but on the other hand working hard and being engaged could also assist in sustaining SA (especially in the case of higher expertise) in which case having a higher MWL can assist in having higher SA (Vidulich & Tsang, 2012). Also having a poor SA does not necessarily have to impinge on performance outcomes, in the case of repetitive, easy tasks, and a poor SA could also cause accumulation of workload which would require SA to increase again (Vidulich & Tsang, 2012).MWL can also be differentiated from attention, in that attentional

resources are fuel allowing the task to be completed and mental workload is the object that the fuel must transport to get the task status from impending to completed (Vidulich & Tsang, 2012). Likewise, the metrics that represent MWL best differ from those that are best for situational awareness and attention, because while all three constructs correlate to an extent, they are distinct. Because of the distinct, yet interrelated nature of situational awareness and mental workload, investigation of mental workload as a viable design assessment tool through both its reliability and whether it can be alleviated for operators through better design is necessary to more human centered control room design.

2.2.5 Human Error

Human error is a key human factors concept that assesses the human element in complex systems, such as control rooms, to investigate how dangerous events occur and what can be done to avoid them (DosSantos et al., 2008). In the human error community, reliability models generally aim to avoid mistakes, which naturally results in a requirement of the system to direct or control the fluctuation of human behavior (Reiman & Oedewald, 2009). However, in the case of human error models it is important that the parameters for what constitutes normal human behavior doesn't become overly restrictive, given that the majority of the time most workers have the best interest of their occupation in mind and will work towards optimizing outcomes and processes (Reiman & Oedewald, 2009). In control room settings human reliability analyses (HRA) are predominantly used for the quantitative assessment of the reliability of human operators and utilize hierarchical task analyses to split a set of tasks into subtasks in order to analyze the most fine-grain components of the system (Pouya et al., 2017). While human error is a prevalent human factors construct, it has been proposed that human error would be somewhat negated by more human centric design (Czaja & Nair, 2012). Furthermore, it has been emphasized that the intricate relationship between humans, technology and organizational factors such as regulatory and managerial processes need to be adequately considered when appraising human error in a given context (Reiman & Oedewald, 2009).

2.3 A Conceptual Framework for Control Room Design

ProRail is the Dutch railway infrastructure manager and is tasked with handling the network railway infrastructure activities, railway traffic flow and allocation and is the employer of the vast majority of TTC's in the Netherlands. ProRail has thus previously investigated a range of prototypes to promote railway innovation in the Netherlands. The future workplace of a TTC has been explored by developing a concept of a machine man interface (MMI) that was based on the current system functionality and the current tasks of a TTC, with the workplace of TTC's in Zwolle being used as a pilot (Lo et al., 2017). A diverse range of materials, methods and techniques were used to determine how the ideal workplace could look, and it was found that intuitive interaction with the task environment alongside precise and clean delivery of information and ergonomic aspects of physical space were all integral factors to be addressed (Lo et al., 2017). In particular future development that is centered around user friendly design was emphasized, with this phase focusing primarily on envisioning the future workplace and exploring ergonomic elements (Lo et al., 2017). Following this research two rounds of a total of six varying hardware setups were assessed (with applications and hardware making the demo up as a whole), and an additional touchscreen which is not present in the current TTC workplace. This phase utilized task analysis,

interviews and eye-tracking to determine what the use of the current system is and explored guidelines that had to be fulfilled by the future workplace, and resulted in three sketch mockups which bigger screens within the ergonomic guidelines were ideal (Lo et al., 2017). Thus, an interface that can not only support cognitive processes, but enhance SA and alleviate the burden of a hefty MW is thus an ideal outcome of usability testing in both control room simulation and mockup (Vidulich, 2012). However, as is the case with many research techniques, usability investigation in the safety critical domain is evolving and fluctuating and there have been contradictions and theoretical gaps in the investigation of usability in safety critical domains and the implementation of human factors methods to improve it.

Given all this, the components of this chapter detail both the control room environment and design process show control rooms to be dynamic, systematic environments that involve operator/system interaction, in a high responsibility organizational context to complete tasks effectively in a time efficient manner. The trend of multiple sociotechnical domains working towards achieving human machine symbiosis imply changes to system design which in turn implies changes to interface design. There is at present a need to introduce HF research early in the design process, rather than devoting resources to solely usability-based research, as well as a need to develop usability research so that it involves quantitative and qualitative measures to enrich the decision-making process for design related outcomes.

These HF concepts and constructs, alongside the UX concept of usability and the focus on quantitative tooling and measurement form the core structure of the current conceptual framework and will be the basis for which the case study in the next section is both designed and operationalized.

Fig 1. Conceptual Framework for Control Room (CR) design

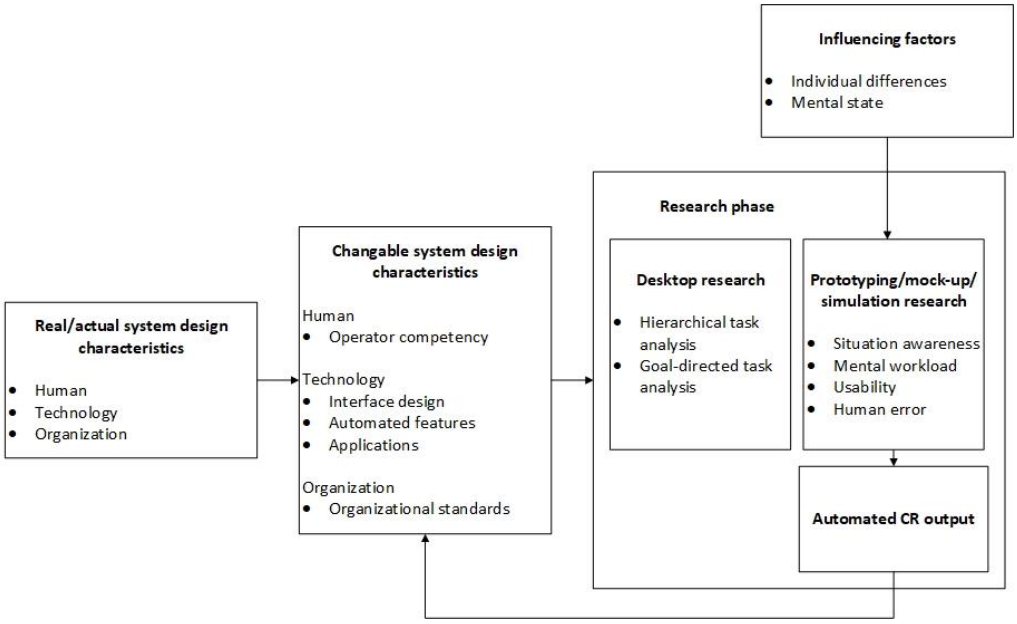


Fig 1. CR design process grouped into three segments, with factors and characteristics of the process clustered into relevant groupings

2.3.1. Actual system design characteristics are intended to depict the broad factors in the control room environment which can be thought of as the pillars of the entire testing process and are steadfast prior to testing. These characteristics represent the limitations of the real-world settings on the design process and give a more well-rounded perspective of the control room environment, and the external environment (organizational and societal) in which it is embedded. Given that high risk organizations are best represented as dynamic systems through a human, technology, organization lens (Karltun & Karltun, 2019), the current conceptual framework has utilized this model to provide a global perspective of the three overarching constraints within which Railway control rooms operate.

2.3.2 Changeable system design characteristics are the many elements that are possible to alter and test that can be grouped under the three overarching constraints.
These elements are intended to be conceptualized as characteristics that are prone to fluctuation or likely to change over time. For example, this can be a particular iteration of a prototype of a system interface, and its specific characteristics such as visual features or functional features. It can also be the manner of, and extent to which technology is implemented in each railway control room. Also, it could be the operator competency or individual operator aptitude sets at a given point in time or in each testing session, or through a given measure. Or it could be the organizational standards, which play a role in research outcomes, but are changeable over time.

2.3.3. The Research Phase
This section represents all the variables to be considered in the development of a control room usability analysis method, and details how to (quantitatively and qualitatively) measure a new interface design in a high responsibility organization context. The research phase is subdivided into a series of groupings, each of which categorizes measures and methods that are used to probe different outcomes and phases in the control room usability analysis process. Suggested measures are discussed in more detail in the following chapter.

2.3.3.1 Desktop Research
Desktop research is the research that is intended to qualitatively delve into the processes operators undergo whilst navigating the system and provide insight into the experience of the operators. Desktop research can be thought of as both research that outlines what is to be expected when operators navigate a given system or simulation, as well as what could be expected given specific qualitative insights.

2.3.3.2 Mockup and Simulation Research
Mock-up and simulation research incorporates human factors constructs into the framework to give a well-rounded representation of what is occurring quantitatively when an operator navigates a novel system environment. The methods and measures in this grouping are intended to give the usability investigation more grounding in terms of objectivity and accuracy, as well as to make results more generalizable and thus promote the unification of some design methods in control room setups, making it a particularly significant aspect of the conceptual framework given that these testing methods are still somewhat preliminary in a usability

investigation context. Influencing factors to the mockup / simulation stage which might represent confounding factors during testing are accounted for, with variable factors such as operator mental state or technological difficulties such as bugs during singular simulations.

2.3.3.3 Automated control room output

The automated control room output represents the culmination of all the other conceptual framework elements into a multidimensional output and is intended to show what the future state of control room assessment could be. Automated CR output is conceptualized as being like the patterned based GUI testing techniques mentioned in 2.2.2.3, but with more complex analysis tools integrated into it (such as, for example, eye tracking), and with an output more specific to the control room context in which the research is being conducted. The output would consist of a range of validated tools and measuring and analysis techniques (for which parameters can be adjusted according to the experimental design at hand), through which data can be fed and recorded, to both generate and output analysis that provides insight into specific design features.

2.3.4. Influencing Factors

The influencing factors that feed into the research phase is intended to portray the possible fluctuating variables which could vary on a given day or in each research environment and influence the outcomes from prototyping or mockup or simulation research.

3. Train Traffic Control ERTMS Simulation Study

3.1 Introduction

The conceptual framework in the previous chapter was constructed to demonstrate the process of control room system testing and design, and the ways in which human factors methods could be implemented to improve the design process. Specifically, it is thought that the introduction of human centric design methods would greatly facilitate the development of more safe and effective control room design and assist in moving away from purely technocentric design (Toft & Kift, 2013; Kovedsi et al., 2018). The eventual goal of automated usability testing based specifically on the procedures and environments used by TTCs in railway control rooms was emphasized in the conceptual framework. It is the intention of the case study presented in this chapter to address a subsection of the conceptual framework and utilize a range of different tooling pertinent to this subsection. The current case study can be thought of as a steppingstone towards future control room design based upon the conceptual framework, by assessing a small cross section of TTCs who work in the ProRail environment. This chapter incorporates a subsection of the conceptual framework by investigating situation awareness using a mixture of qualitative and quantitative tooling in the context of new and old interface comparison.

The current case study was conducted in the context of the European Rail Traffic Management System (ERTMS) during the Covid 19 period. ERTMS is currently undergoing widespread testing and implementation throughout the EU and will be implemented as a technology in coming years by the many socio technical organizations, such as ProRail. ERTMS aims to unify international EU train travel systems by using in-cab and trackside devices to enforce a standard for safety by supervising speed and using European-wide standard signaling systems. Additionally, ERTMS is also intended to enhance efficiency and cross-border train traffic flow throughout the entire EU, making it a forward focused context to appraise the workplace and way TTC's will be working in the future. ERTMS is currently used in the Netherlands on HSL lines (a high-speed line between Amsterdam with the Thalys train) and is also going to be more widely used (for example between Rotterdam port and Germany). Because TTCs will work with an ERTMS area and a regular area, the case study was used to investigate the ERTMS simulator interface compared to the current workplace interface, in terms of both usability and SA enhancement. Given that scenario one was designed to be simpler than scenario two, and that many features of the ERTMS interface were novel, it was hypothesized that participants would perform better in terms of all metrics and in the directions of what would indicate elevated SA levels in scenario one as opposed to scenario two, and in the current interface as opposed to the ERTMS interface.

3.2. Method

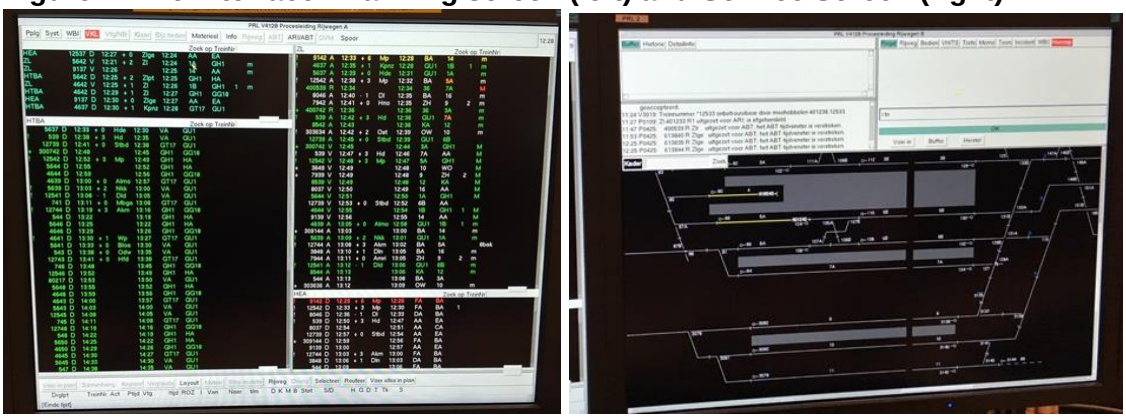
3.2.1. Experimental design

3.2.1.2 Simulator

The current experiment utilized a simulator with two interfaces, of which one represented the current interface, and one represented the future ERTMS interface. The simulator consists of an interface with tracks located in Oudenbosch, which is in the vicinity of North Brabant in the southern region of the Netherlands. The simulation setup for both interfaces consisted of four

screens, with the left screen containing the planning screen, the middle the service area with a service menu at the top, and the right two screens containing an overview of the geographical area (called an oversight screen) which the operator can manipulate to gain insight into what is occurring, and a zoomed in area of the oversight screen where scenarios were occurring. Both interfaces had a dark screen background, with grey lines representing tracks, and rectangles with numbers representing trains. The new European Railway Traffic Management System (ERTMS) interface features facilitated TTCs having access to more information surrounding the trains, such as train mode, and specific speed and location. Also, TTCs were given more flexibility in controlling features such as implementation of speed or speed limitations in either track sections, or the entire track, with the possibility to fill out reasons for the implementation. Likewise, the idea of the new interface was to support a more extensive shared mental model between the TTC and driver, by giving both parties access to the same information, which allowed for more efficient communication. Because the ERTMS interface was used to introduce a new visual system in which different color schemes are used to represent different things (such as a feature that shows colored tracks indicating different levels of control drivers have) it had generally more color than the current interface.

Figure 2. The Interface: Planning screen (left) and Service Screen (right).



3.2.1.3. Conditions

Four conditions were made for the current experiment. Two scenarios of varying complexity were developed (scenarios one and two), both of which were presented on the simulator in both the current interface and the ERTMS interface, which made up the total four conditions (see table one below for an overview). Scenario one was split into a training segment at the start of fourteen minutes, as well as a normal service (which consisted of shunting and monitoring activities) and a light disturbance segment which consisted of the additional two tasks of giving instructions in abnormal situations and entering reports into the interface. The training segment allowed operators to acquaint themselves with the new service area in the current interface (since the service area differed from the one, they worked with daily), and to acquaint themselves with the ERTMS functionalities in the ERTMS interface. Scenario two required TTCs to complete all the previous tasks in addition to coping with a large disturbance and more delays, making it a medium to heavy workload level. All conditions took a total of twenty-eight minutes to complete from the time they started. Both conditions with scenario one always ran in the order of current interface followed by ERTMS interface, to minimize the effect of the new service area on performance by allowing exposure to the new service area before introducing

ERTMS. The order of scenario two was shuffled per participant to minimize confounding variables.

Table 1. Conditions

	A. Current UI	B. ERTMS UI
Scenario 1. (Light to medium delays)	1a	1b
Scenario 2. (Heavy delays, disturbance to service)	2a	2b

The scenarios had SA probe queries occurring every seven minutes to get a dynamic snapshot TTC SA. Because scenario one consisted of a training period at the start, there were only two SA probe pauses with a total of six questions per pause, to maximize time efficiency. Scenario two had three pauses with a total of four questions per pause. In total, per condition, twelve probe questions were administered. The scenario structure and the embedded probe query structure represented a trade-off for each other, in that both had to be adapted meticulously for complex events to occur how they would in the field, whilst ensuring that probe queries could occur punctually to ensure rigorous experimental design.

3.2.2. Participants

Nine fully trained TTC's, with previous ERTMS experience (relating to a dual signaling area) working at the Zwolle train traffic control center participated voluntarily in the current study. The participants had on average M^1= 12.5 (SD^2 = 10.5) years of experience as TTCs, and M = 13.5, SD = 11.5 years of experience in the rail sector.

3.2.3. Metrics

As detailed in the conceptual framework, the integration of highly complex systems (such as the control rooms where TTCs work) usually results in an overload of information conveyance through channels, which is where human error and other human factors constructs come into play (Endsley, 2012). In the case of high responsibility domains, it is important to recognize that true SA exists within the operators own mental model and design should thus be centered around this (Endsley, 2012). SA will thus be the linchpin for the experimental design of the current case study, with the setup exploring tools for SA measurement.

3.2.2.1. Situation Awareness Eye Tracking Measures

Eye-tracking measures have been pivotal in the development of quantitative assessment of interface environments and have been used in the human factors research field in conjunction with subjective HF construct measuring tools for the most part, as well as probe queries to a lesser extent (see Table 1 for a review in the appendix). However the distinction between which eye-tracking metrics translate more to mental workload and which ones translate more to situation awareness has been oftentimes unclear, given that many experiments investigating this phenomenon have used either one subjective measuring tool to associate the construct level with the eye-tracking metric, or a probe query, or another type of physiological measure, and rarely have all four types of metrics been integrated together, with both MWL and SA subjective measures as a comparison. Zhang et al. (2020) conducted a systematic review that concurred with this notion, but found that overall, more conscious, and goal-oriented eye movements (such as fixation, and dwell metrics) had a stronger relationship with SA measures than more unconscious eye movements (such as blink and pupil metrics). Given that there is no

precise standard at present the current study will incorporate measures that have been traditionally thought to be associated with both MWL and SA, to tease out which eye-tracking metrics represent SA and to what extent. See table three in the appendix for a broad overview of eye-tracking metrics which are deemed to be pertinent to quantitative human factors assessment of control room design, and their definitions and predominant usage.

Several more conscious eye tracking fixation measures (fixation duration, rate, and proportion) were utilized in the current experiment to assess to what extent they represent SA, which is in line with previous research (see tables one and two in the appendix). Fixation duration is conceptualized as prolonged focus of attention (that is when the gaze remains fixed), pinpointed to a specific area, and has previously been found to represent adaptation to novel situations, as well as significant association to self-report SA and MWL measuring tools (Ikuma et al., 2014; Petersen et al., 2019). Fixation proportion and rate, both of which are numerosity measures and will be measured per trial (during the parts where the scenario is running, and not when the probe queries occur), and have been shown to be negatively related to level of experience, as well as performance (Argyle et al., 2020; Holmquist & Nystrom, 2011; Coyne & Sibley, 2015; Bracken et al., 2019, Di Flumeri et al., 2018). The current experiment also investigated a range of less conscious, saccade measures (saccadic velocity, saccadic amplitude, and saccadic duration) to assess to what extent less conscious eye tracking metrics can account for SA. Saccadic velocity is calculated using the average velocities of saccades occurring during a specific time window and has previously been thought to be inversely associated with target predictability (Holmquist & Nystrom, 2011), making it pertinent to comparison between the interfaces. Saccadic amplitude is conceptualized as the displacement of gaze, in pixels, and has previously been thought to be inversely associated with focused inspection of an environment (Tatler & Vincent, 2008; Recarte & Nunes, 2003). Saccadic duration is defined as the time which passes between two smooth pursuits (which is when the eye remains fixed on a moving target) or fixations when the eye remains fixed within a specified area for a specified period), and were measured in the current experiment in order to assess the nature of their relationship with SA since it has previously been studied with MWL only, and found to have a negative association to increased levels of MWL (Marandi et al., 2018; Holmquist & Nystrom, 2011). The saccadic and fixation events were processed using a robust event classification algorithm, REModNaV, which uses velocity-based eye movement event classification, and is based on the algorithm of Nyström and Holmquist (2010). REModNaV defines fixations as a time period in which a stimulus is focused upon, with the minimal threshold for a fixation duration being 0.04 seconds (or 40 milliseconds), with fixations being events that are higher than this threshold value and do not qualify as pursuit or saccade events (Dar, Wagner & Hanke, 2020). Saccades are conceptualized by REModNAV as moments in which eye movement occurs towards a new target region and visual intake is subdued, with the minimum duration for a saccade qualifying as 0.01seconds (or ten milliseconds), and the minimal inter saccadic duration being 0.04 seconds (Dar, Wagner & Hanke, 2020). These values and definitions are in line with previous validated research done by Nyström and Holmquist (2010).

3.2.2.2. Situation Awareness Probe Queries

Conventionally situational awareness has been measured in control rooms with explicit

measures such as the situational awareness global assessment (SAGAT), which is a query used in real time, human in the loop simulations in design evaluation contexts (Endsley et al., 2000). The SAGAT responses correspond to the SA three levels (perception, comprehension, and projection), and are attained by freezing the interface at random moments and querying operators about their perception (Endsley et al., 2000). Another similar measurement called the situation present assessment method (SPAM) can be differentiated from the SAGAT because it uses response delay as the dependent variable to assess vigilance and was developed with the rationale that it isn't obtrusive because the task and interface aren't frozen (Bakdash et al., 2020). SPAM is advantageous because it dynamically assesses the SA while the situation is occurring by using the response time (Miramontes, 2017). Recently a meta-analysis on the two query types found both to be equally predictive of performance, but found that SPAM resulted in problems with intrusiveness, but that the SAGAT (while overall was found to be more sensitive and reliable) had problems with memory reliance (Endsley, 2019). However, it is also arguable that SAGAT is less appropriate to measure something as dynamic as SA since the operator is out of the loop and the display is frozen (Nachtwei, 2016). Moreover, it has been found that TTC's in the Netherlands have low SAGAT levels but high performance on a series of performance measures in simulation experiments, making it an inaccurate tool for SA analysis in the TTC control room (Lo et al., 2017). It is theorized that the discrepancy in outcomes is related to implicit processes that are conditioned through experience (Lo et al., 2017). On the other hand, since SPAM creates additional workload by adding a secondary task, and SAGAT likely reduces workload by taking the operator out of the loop, it is important to consider the implications of this with regards to assessing workload in control room settings (Nachtwei, 2016). More recently it was found that SAGAT resulted in higher workload in simulation settings than SPAM, and that query outcomes were better for SPAM queries than SAGAT queries (Fujino et al., 2020). Given that SAGAT has already been trialed in a similar environment (Lo et al., 2016) and is associated with poorer query response, it was decided that an adapted SPAM with questions corresponding to the three levels of SA would be used in the current experiment. For an overview of previous implementation of SPAM, please refer to table two in the appendix.

3.2.2.3 Situation Awareness Self Rating Measures

Self-rating measures provide insight into the subjective rating the participants have of themselves for a particular construct. In the case of SA, the most widely known measures used are the Situation Awareness Rating Technique, which has had widespread use in conjunction with experiments which have aimed to assess eye tracking measures and SA (see table one in the appendix), and uses ten dimensions to assess SA (Salmon, 2006). The Mission Awareness Rating Scale comprises questions based upon the three level SA model (Endsley, 1995) and is administered after mission completion. Although MARS elicit responses which correspond to the Endsley (1995) three levels of SA, SART has been much more frequently applied to studies which also utilized SAGAT and eye tracking. The current experiment utilized MARs because of this, since the SPAM probe queries were adjusted to correspond to the three levels of SA (identification, comprehension and projection), and also since it was developed for applied research settings (Salmon, 2006).

3.2.4. Materials

3.2.4.1 Eye-tracker

The eye tracker used in the current case study was an immobile eye tracker from Sweden, the Smart Eye Pro Eye-tracker with version 9.1 software. Five lenses were positioned symmetrically, and adjusted for resolution and brightness, and a coordinate system was constructed through a series of calibration procedures to create a tracking space within the remote eye tracker setup. A combination of lenses of 8mm and 6mm were used in order to attain the best accuracy of data without compromising the headbox space, which could have ramifications for both data quality and operator comfort.

3.2.4.2. Hardware

Four Dell U2515H monitors were used to display the simulation of both interfaces and were set to 1920 x 1080 resolution.

3.2.4.3. Questionnaires:

i) Probe Queries

The probe queries presented at intervals throughout the simulation consisted of both the SPAM queries, with response times and accuracy of responses as SA measures, and SA and MWL self-report items which were presented following this. An adapted version of Durso et al. 's (1999) Situation Present Assessment Method was used in order to assess SA dynamically, in the moment at which it was occurring. The probe itself was presented in the form of a popup icon that would appear once participants clicked on another icon. The questions were created with the support of two subject matter experts (TTC's) and were modelled from what was occurring in the situation. Each freeze probe contained either four or six multiple choice questions, of which there were always questions addressing Endley's (1999) levels one, two and three situation awareness states. The intention was to spread the different levels as evenly as possible across different points in the scenarios to provide an accurate cross-section of TTCs' SA, but due to the nature of the content, it wasn't always possible to have an even number of questions for each level in each scenario. However, there was no strict guide that dictated that levels of questions had to be spread totally evenly across scenarios in terms of levels, since SPAM doesn't normally address levels of SA, making it justifiable to incorporate dynamic SA level testing.

The questionnaire items presented following the SPAM queries consisted of four items from Matthew and Beals (2002)'s Mission Awareness Rating Scale, a subjective self-rating scale that has been found to be significantly related to performance outcomes in safety critical environments (Hong et al., 2015; Matthews, 2011; Salmon, 2008), as well the integrated workload scale, which has been shown to be a verified subjective mental workload tool that has been previously applied in Dutch railway settings and captures the multidimensional nature of train traffic controller MWL with items addressing effort, demand and time (Kramer et al., 2016).

ii) Debriefing questionnaire

A debriefing questionnaire was created to address the mental state of the TTCs post testing, as well as collect subjective measures and validity measures. This can be found with the at the end of the appendix. The questionnaire items served the purpose of controlling whether the

simulator was more mentally demanding (due to the difference in service area), as well as addressing how quickly TTCs could adapt to the simulator and how valid it was in terms of how TTCs work day to day. Additionally, two questions were posed to address operator experience, their ratings for additional worth of the simulator and a box for suggestions for the sort of educational material they might like to have to prepare for an ERTMS workplace in the future. A further ten questions based on Lankveld et al. (2017)'s study on Gaming Simulation Validity addressed the statistical validity of the simulator with items relating to process validity, Psychological reality and structural validity. Usability was assessed with another thirteen questions using the System Usability Scale, which has been found to be reliable in applied research settings (Peres et al., 2013; Kaya et al., 2019). Likewise, a ranking question was included, in which TTCs ranked various ERTMS functionalities in terms of usefulness in order to allow for collection of further qualitative insight. At the end of the questionnaire there was a question for qualitative insights into the experiment, and operators also participated in a think aloud debriefing with the simulator operator and the researchers. All questions were translated by a native Dutch speaker with experience in the matter or used from previous validated research in which the questions had been posed in Dutch.

3.2.5. Procedure

Prior to the experiment participants were given an instructional PowerPoint which both detailed what to expect in terms of the simulator content and visual representation of ERTMS and served the purpose of refreshing operator ERTMS knowledge prior to the actual experiment in order to alleviate learning workload that would occur throughout the scenario. Likewise, TTCs had access to the ERTMS instruction module via E-learning. On the test days themselves there was an initial introduction period, during which the aims of the research, the testing schedule, and the impact the research would have on the development of ERTMS were explained. There was also a series of instructions given to the TTCs partaking in the study about the way the eye-tracker worked, how their privacy would be protected, and how it was imperative that they move as little as possible to allow for accurate data collection. Likewise, the way the probe queries would work was clarified. Following this the eye-tracker was calibrated for each TTC, and the four scenarios took place for a duration of three hours, during which the probe queries occurred. Separate eye tracking recordings were taken for each condition to minimize the amount of excess data collected, and a roughly twenty-minute break was taken between scenarios. At the end of which participants filled in the questionnaire. Three participants did not complete one of the conditions, due to time restraints with the scheduling (this is accounted for in the results). The current experiment also utilized informal current and retrospective think aloud (in the form of a debriefing), with educated moderators collecting and processing the qualitative data. Likewise, the first ten-minute interval of scenarios 1a and 1b incorporated somewhat open-ended introductory phases to allow the TTCs to become both acquainted with the system and give preliminary feedback from a novel perspective. Following the simulation all collected data was anonymized prior to analysis, and the eye tracker data that was recorded during the probe queries in each condition was removed (since the scenario wasn't occurring during these periods). The SPAM data was analyzed by the researchers and subject matter experts to determine how accurate each of the TTCs was in scoring.

3.3. Results

3.3.1. Eye Tracking Results

Prior to assessing eye-tracking measures, the data per scenario was filtered so that the time during which SPAM probe queries were occurring was filtered out (since the scenario wasn't occurring during this period and this would confound SA findings). Table 2. below depicts the eye-tracking metric averages per condition. Fixation rate and duration were higher in the current interface than the ERTMS interface in scenario one, but lower (or equivalent in the case of rate) in scenario two. Saccadic measures generally didn't show surface level patterns when eye-balling the means and standard deviations. To assess whether participants performed significantly differently on the same eye tracking measures between conditions paired samples t-tests were conducted for each measure between each condition. Fixation duration averages were found to be significantly higher in the ERTMS interface than the current interface in scenario 2 ($t(6)=-2.76$, $p=0.03$*), as well as being significantly higher in the current interface in the more difficult scenario two than in the easier scenario one ($t(5)=2.46$, $p=0.05$*). Moreover, average saccadic velocity was found to be significantly higher in scenario two than scenario one in the current interface ($t(5)=-5.58$, $p=0.003$*) and significantly higher in the ERTMS interface in scenario one ($t(7)=2.416$, $p=0.046$*).

Table 2. Means and standard deviations for eye-tracking measures, per scenario and interface.

		Current UI		**ERTMS UI**	
		M	**SD**	**M**	**SD**
Scenario 1	Fixation Duration (ms)	150	5	140	3
	Fixation Rate (per min.)	22.92	4.28	21.80	4.94
	Saccadic Duration (ms)	31	0.5	29.5	0.5
	Saccadic Velocity (°.sec)	0.15	0.01	0.17	0.02
	Saccadic Amplitude (°)	0.003	0.0005	0.008	0.016
Scenario 2	Fixation Duration(ms)	120	3.6	140	4.5
	Fixation Rate (per min.)	25.60	5.41	25.60	3.43
	Saccadic Duration (ms)	30	0.4	40	0.1
	Saccadic Velocity (°.sec)	0.17	0.007	0.16	0.014
	Saccadic Amplitude (°)	0.002	0.000	0.003	0.000

3.3.2. Probe Query (SPAM and Reaction Time) and Self Report (MARS) Results

Depicted in table three below are the average SPAM scores, reaction times and experienced SA both per scenario and per interface. A marked difference is noticeable between scores for both interfaces and scenarios, with a trend of scores being higher in scenario two for both interfaces, and the ERTMS interface for both scenarios. Paired samples t-tests demonstrated significantly better reaction times ($t(7) =2.9$, $p = 0.02$) and significantly higher self-rated SA on

average ($t(7) = 2.9$, $p = 0.02$) for the ERTMS UI in scenario one. On the other hand, paired samples t-tests conducted on scenario two results showed that only the SPAM scores differed significantly ($t(6) = 2.63$, $p = 0.04$.), although overall patterns were the same, with ERTMS UI again having a better score, lower reaction times, and higher self-rated SA.

Table 3. Mean SA scores, relevant reaction times and experienced SA.

		Current UI		ERTMS UI	
		M	**SD**	**M**	**SD**
Scenario 1	SA score (%)	67.9	25.8	84.4	16.6
	Reaction Time SA score	32.4	8.9	21.3	9.7
	Experienced SA (1-4)	3.1	0.3	3.3	0.3
Scenario 2	SA score (%)	72.3	17.6	91.8	10.6
	Reaction Time SA score	23.9	0.2	21.6	0.2
	Experienced SA (1-4)	3.1	7.9	3.4	7.5

Table four depicts the average scores per SA level for SPAM questions and their respective reaction times and self-rated SA and shows consistently higher scores for the ERTMS UI with regards to the SA scores, reaction times and SA self-ratings. A striking result is that the level two scores are higher than the level one scores in scenario one, and the level three scores are higher than the level two scores in scenario two. According to the SA model, level one scores should be the highest since perception is the basest level of the model. However only the difference between levels two and three in scenario two was significantly different when looking at paired samples Wilcoxon tests (t=0.00, p=0.022.)

Table 4. Per SA level: SA scores, Reaction times and self-rated SA.

		Current UI		ERTMS UI	
		M	**SD**	**M**	**SD**
Scenario 1	Level 1 SA score (%)	64.6	35.4	86.8	21.9
	Level 1 Reaction-time score	29.9	6.6	22.4	10.3
	Level 1 Self-rated SA (1-4)	2.4	0.5	3.4	0.3
	Level 2 SA score (%)	62.5	17.6	89.0	14.0
	Level 2 Reaction-time score	38.0	11.9	26.1	7.7
	Level 2 Self-rated SA (1-4)	3.1	0.2	3.4	0.1
	Level 3 SA score (%)	56.3	8.8	67	13.8
	Level 3 Reaction-time score	27.8	4.3	22.8	8.9

	Level 3 Self-rated SA (1-4)	3.2	0.1	2.9	0.0
Scenario 2	Level 1 SA score (%)	76.0	14.3	96.3	6.4
	Level 1 Reaction-time score	22.2	3.9	14.9	1.4
	Level 1 Self-rated SA (1-4)	3.1	0.0	3.4	0.2
	Level 2 SA score (%)	57	6.4	89.0	5.7
	Level 2 Reaction-time score	26.5	11.7	22.0	5.7
	Level 2 Self-rated SA (1-4)	3.2	0.3	3.4	0.3
	Level 3 SA score (%)	82.4	19.1	91.2	9.0
	Level 3 Reaction-time score	23.3	5.5	24.4	10.4
	Level 3 Self-rated SA (1-4)	3.2	0.2	3.3	0.1

Correlations between the three SA variables, reaction time, SPAM accuracy score and self-rated SA for the entire experiment went in the expected directions when looking at Pearson correlation coefficients, with non-significant negative correlation between reaction time and both SPAM and MARs results (r=-0.20, p=0.61 and r=-0.41, p=0.27 respectively), confirming that longer reaction times is an indicator of lowered SA. Also, a significant, high positive correlation was found between the SPAM accuracy and MARS ratings (r=0.75, p=0.021).

3.3.3. SPAM, MARS and Eye Tracking results

Given that previous research has shown Pearson correlation coefficients to be an appropriate measure for assessing the nature of the relationship between eye tracking metrics and both exploratory SA measures and subjective measures (De Winter et al., 2019; Hasanzadeh et al., 2016), and given the small sample size of the current experiment, the overall relationship between SA metrics and eye tracking metrics was assessed using Pearson correlation coefficients. These were done between each eye tracking metrics and each SA metric for the entire experiment, as well as per interface and per scenario, between each eye-tracking metric and each SA metrics. Overall, the only significant correlations when looking at Pearson correlation coefficients were exhibited between reaction time, and both fixation duration (r=-0.83, p=0.00) and fixation rate (r=0.82, p=0.01).

When appraising table five (below), scenario one in the current interface (that is condition 1a) showed the highest number of low associations between the various eye-tracking measures and the probe query and self-rate SA measures. Scenario two in the current interface (2a) had the most moderate to high associations (between 0.5 and 0.75) and the highest (0.75-1) associations, with a high associated value between the fixation duration and the SPAM score, and a high association between aggregate saccades and the reaction time. Significant high associations were also found in 2a between saccadic duration with both the reaction time, and

self-rated MARS scores, as well as between aggregate saccades and the self-rated MARS score (see table 4. below). In the same line, two significant associations were also found in scenario 2b (the more difficult scenario with the ERTMS interface) between saccadic amplitude and aggregate saccades and reaction time, as well as a negative significant correlation between the MARS scores and aggregate fixations.

Saccadic velocity was the only eye tracking measure with no significant association to the other measure types in all four conditions and was composed of half moderate to high associations and half low ones. In terms of positive and negative associations, fixation duration was negatively correlated (mostly moderately) with reaction time and MARS scores in all cases but one throughout the four conditions, as well as being negatively correlated with SPAM in scenario two in both interfaces. In three of the four conditions saccadic duration correlated negatively with MARS scores, but only once significantly. Saccadic amplitude had either a low positive or a moderate to strong negative association with reaction time in each condition, with the strong negative association being significant. Overall aggregate fixations had either very weak or strong negative associations to measures, with two significant negative associations with SPAM scores in 1a and MARS score in 2b. Aggregate saccades always correlated positively with SA measures and were moderate to high across all conditions except condition 1a (scenario one, current interface).

Table 5. Correlations between Eye-tracking measures and SA measures per condition.

		Current UI			**ERTMS UI**		
		SPAM score (%)	Reaction Time (s)	MARS score (1-4)	SPAM score (%)	Reaction time (s)	MARS score(1-4)
Scenario 1	Fixation Duration (ms)	r=0.273, p=0.554	r=-0.46 , p=0.297	r=-0.054 , p=0.9	r= 0.10, p=0.79	r=-0.80 , p=0.009*	r=-0.157 , p=0.69
	Fixation Rate (per min.)	r=-0.78 , p=0.04*	r=0.403 , p=0.37	r=-0.32 , p=0.48	r=0.225 , p=0.56	r=0.55 , p=0.12	r=0.44 , p=0.24
	Saccadic Duration (ms)	r=-0.52 , p=0.23	r=0.57, p=0.214	r=-0.19 , p=0.68	r=0.15 , p=0.70	r=-0.5 , p=0.16	r=0.35 , p=0.35
	Saccadic Velocity (°.sec)	r=0.17 , p=0.71	r=0.19 , p=0.69	r=-0.32 , p=0.49	r=0.13 , p=0.74	r=0.41 , p=0.28	r=0.41 , p=0.27
	Saccadic Amplitude (°)	r=-0.311 , p=0.497	r=0.20 , p=0.62	r=-0.045 , p=0.923	r=0.31 , p=0.41	r=-0.32 , p=0.40	r=-0.28 , p=0.46
	Aggregate Fixations	r=-0.73 , p=0.05*	r=0.136 , p=0.771	r=-0.297 , p= 0.51	r=-0.48 , p=0.92	r=0.22 , p=0.57	r=0.09 , p=0.83
	Aggregate Saccades	r=0.00 , p= 0.99	r=0.118 , p=0.8	r=-0.17 , p=0.71	r=0.44 , p= 0.24	r=0.49 , p=0.18	r=0.61 , p= 0.08*
Scenario 2	Fixation Duration(ms)	r=-0.90 , p=0.84	r=-0.50 , p=0.26	r=0.66 , p=0.11	r=-0.32 , p=0.41	r=-0.29 , p=0.45	r=-0.33 , p=0.40

Fixation Rate (per min.)	r=-0.05 , p=0.91	r=0.15 , p=0.75	r=-0.58 , p=0.17	r=0.56 , p=0.19	r=0.29 , p=0.45	r=-0.33 , p=0.38
Saccadic Duration (ms)	r=0.28 , p=0.544	r=0.91 , p=0.004*	r=-0.83 , p=0.002	r=0.33 , p=0.38	r=0.15 , p=0.71	r=-0.46 , p=0.21
Saccadic Velocity (°.sec)	r=-0.25 , p=0.59	r=0.50 , p=0.26	r=-0.44 , p=0.32	r=0.02 , p=0.96	r=0.44 , p=0.24	r= 0.22, p=0.57
Saccadic Amplitude (°)	r=0.174 , p=0.71	r=0.28 , p=0.54	r=-0.22 , p=0.63	r=-0.41 , p=0.27	r=-0.69 , p=0.042*	r=-0.59 , p=0.09
Aggregate Fixations	r=-0.02 , p=0.97	r=-0.09 , p=0.85	r=-0.24 , p=0.61	r=0.216 , p=0.56	r=-0.24 , p=0.531	r=-0.73 , p=0.03*
Aggregate Saccades	r=-0.05 , p=0.92	r=0.71 , p= 0.071	r=-0.76 , p=0.05*	r=0.66 , p= 0.05*	r=0.84 , p= 0.005*	r=0.65 , p= 0.05*

*Significant findings are detailed with an **

3.3.5. Questionnaire Results

Participants rated their ERTMS knowledge as relatively good with regards to completing the scenarios well (*M* = 3.5*, SD* = 0.9), and that they could acquaint themselves relatively quickly to a different service area (*M* =3.6, *SD* = 0.9), but it was ascertained during the debriefing that they experienced a higher mental workload due to getting used to a new service area. With regards to the simulator, participants could acquaint themselves quickly to the environment (*M* = 3.8, *SD* = 0.2.), and experienced the situation as relatively realistic (*M* = 4.4*, SD* = 0.5.). Also, the participants found the second scenario as easier, because of completing the first scenario (*M* = 4.4*, SD* = 0.5), as well as finding the instruction material adequate as a preparation tool for the simulation (*M* = 4.2, *SD* = 0.7). For the ERTMS usability analysis, an overall relatively high positive rating was found once negated statements were converted (*M* = 4.0*, SD* = 1.1). However, a striking finding was the relatively low trust score, for which participants still feel quite unfamiliar with the ERTMS interface (*M* = 2.6, *SD* = 2).

3.4. Discussion

The aim of the current thesis was to firstly contribute to control room design analysis by exploring relevant human factors techniques and processes pertinent to a human centric design approach and control room control room design analysis, and to integrate this information into a conceptual framework for future use by human factors professionals in the context of control rooms. Secondly the current thesis aimed to explore a subsection of the conceptual framework through the current experiment, which used human factors quantitative research methods to compare two interfaces, in the context of control room railway design. The plausibility of integrating quantitative human factors research techniques into design methods in control rooms was investigated. It was hypothesized that the preliminary findings from the case study would support the notion that more extensive use of quantitative human factors research methods in control room design can add more dimensionality to control room design and usability research. Furthermore, it was also hypothesized that participants would perform better both in terms of reaction times and SPAM outcomes in conditions with scenario one, and in the current interface as opposed to the ERTMS interface. Results from the current case study

indicate that quantitative human factors techniques such as eye tracking and probe queries are useful in the assessment of new interfaces and train traffic control procedures and contribute to effective design investigation. It was also hypothesized that because scenario one was designed to be simpler than scenario two, and because many features of the ERTMS interface were novel, participants would perform better in terms of all metrics and in the directions of what would indicate elevated SA levels in scenario one as opposed to scenario two, and in the current interface as opposed to the ERTMS interface. This hypothesis was not supported since participants performed overall better in scenario two than scenario one, and presented mixed findings for the variety of measures between the two interfaces, with a slight inclination for better performance in the ERTMS interface.

3.4.1. Main Findings

Overall a general trend was shown amongst fixation eye tracking measures (that is, duration, rates, and aggregates) as being on average lower in scenario two than one. Furthermore, the mean fixation duration values of scenario one in the current interface were found to be significantly higher than in scenario two. Previous findings have demonstrated the plausibility of using fixation duration as an eye tracking measure to represent SA, by investigating it alongside self-rated SA and/ or probe query SA outcomes (Argyle et al., 2020; Bracken et al., 2019; Di Flumeri et al., 2019). In line with this SPAM scores and self-rated SA were both higher in the second scenario than the first, meaning that operators both rated their SA as higher and performed better in this condition, which fits with previous findings. In the current interface this pattern was also represented in reaction times, with longer reaction times in scenario one, which represents a lowered SA. While this is in contrast to the hypothesis of the current experiment that operators would perform better in the easier condition and in the current interface, it is in accordance with the questionnaire finding that operators found it relatively cumbersome to adapt to a new service area during scenario one in both interfaces.

In terms of correlational associations, SPAM accuracy results were found to have a high association with the self-rated MARS results, which lends support to the notion that SPAM can be used to assess situation awareness. This is in line with previous findings that validated self-rating SA tools and SA probe query tools such as SAGAT and SPAM are positively associated and can be utilised in the assessment of SA (Bracken et al., 2019; Strybel et al., 2008; Durso et al., 2006). With regards to overall fixations versus saccadic event outcomes, both fixation measures and saccade eye tracking measures had equal amounts of high correlation outcomes overall when assessing correlations per condition, with equal amounts of significant outcomes. Previously many findings have mainly focused on fixation measures when assessing SA using eye tracking measures (Argyle et al., 2020; Desvergez et al., 2019; Petersen et al., 2019). However, it should be noted that fifteen percent of the correlations done overall, when assessing correlations between eye tracking measures and SA measures per condition, had strong associations and/ or correlated significantly. Saccadic velocity had low to low-moderate associations with the SA measures in all outcomes and all conditions, making it the weakest associated measure. However, saccadic velocity did exhibit significantly different outcomes between its means for scenarios one and two in the current interface and was significantly higher in scenario one for the current interface. This suggests that it should not be disregarded

altogether, but perhaps investigated more extensively when either comparing a new interface to a control one in a learning condition, or a more difficult scenario to an easier one in a control interface.

Overall, more eye tracking events were shown to occur in scenario two than one, and more were shown to occur in the ERTMS interface than the current interface on average. This was an interesting finding, given that overall performance on the three SA measures (MARS, SPAM and SPAM reaction time) followed the pattern of having better outcomes in scenario two and in ERTMS conditions. This was in line with the outcomes shown by correlations between aggregate saccades and the SA measures in all conditions, whereby higher positive (and mostly significant) associations with SA measures were shown in scenario two in both interfaces, and lower non-significant associations were shown in scenario one where SA outcomes were worse. This demonstrates that more saccades occurring, could be an indication of elevated SA. Aggregate fixations on the other hand, had weak associations for the most part in all conditions, except for two significant associations, making it the most ambiguous eye tracking measure for representing SA in this experiment. However, fixation rate and duration had some high, significant, negative correlations which indicates that the number of fixations occurring might be irrelevant, but that length and frequency of these events could indicate SA levels. Since fixation duration and frequency have both been previously found to be significantly related to SA probe query and self-rated SA (Miramonte, 2017; Ikuma et al., 2014; Moore & Gugerty, 2010) it is assumed that a replication of the current study with a larger sample size could produce this result in more conditions.

3.4.1.2. Per Scenario

With regards to the scenarios specifically, the only significant associations in scenario one between SA measures and eye tracking measures were all fixation eye-tracking measures, whereas in scenario two they were a mix of saccade based and fixation-based measures, that tended more to saccades. This could suggest that fixations are more useful in assessing SA in novel situations. However, since a pattern of higher associations was shown in the second scenario for both interfaces, and more events also occurred in this domain, this finding should be interpreted lightly. Moreover, seventy percent of the significant correlations between eye tracking measures and SA measures occurred in the scenario two conditions. This could be due to the learning portion of scenario one, whereby operators received instruction from the Game Leader who was sitting behind them, which could have caused their gaze to wander from the screen, and a loss of data and thus less data to indicate overall gaze behavior during this time period. On the other hand, adaptation in scenarios 1a/1b to the new service area in the simulator or the research environment could be reflected through the weaker associations between eye tracking measures and SA measures, and the lesser number of events occurring in the two scenario one conditions relative to the two scenario two conditions. Saccadic duration was mostly moderate to strong in both scenario two conditions, with significant positive associations to reaction time and MARS in 2a, suggesting that it could be a useful measure for evaluating SA in a control interface which operators are accustomed to, and have experienced a learning phase with. Since saccadic duration has been previously found to be non-significantly associated with MWL (Di Stasi et al., 2010), it is recommended that in the future the association

between saccadic velocity and SA measures as opposed to MWL measures is explored to assess whether it can represent SA better.

3.4.1.3. Per Interface

When appraising the two interfaces, a broad finding was that more events occurred altogether in the ERTMS interface than the current interface, although average reaction times, self-rated MARS SA outcomes and SPAM outcomes were generally better in ERTMS, with the first two being significantly better in scenario one. This finding and the finding that aggregate saccades and SA measures exhibited strong to moderate correlations in the ERTMS conditions encourages the notion that the number of saccades occurring is useful in determining SA. These findings could also be useful in further investigation of the relationship between usability and eye tracking measures since ERTMS usability outcomes were generally high and positive. Furthermore, saccadic amplitude and SA measures correlated negatively and more moderate to strongly in the ERTMS conditions, as well as more highly in both scenario two conditions which suggests future use of saccadic amplitude in testing SA in a new interface, with a higher amplitude suggesting a lowered SA. Previous research has likewise found saccadic amplitude to be a promising SA eye tracking measure (Zu, Coster & De Winter, 2017). These findings imply that saccadic amplitude and aggregate saccades are promising eye-tracking measures for evaluating SA and usability outcomes when operators are navigating a new interface, to compare both more easy and difficult new interface conditions, as well as control interface and new interface conditions once operators have already experienced a sufficient learning period.

3.4.1.4. Scenario two: Comparing ERTMS and Current UIs

Since it is possible that scenario one served the purpose of being a learning period given the weak correlations and lesser number of significant findings, this section will address results between scenarios 2a and 2b specifically. In scenario two both interfaces exhibited the same number of significant correlations, but between different eye tracking measures and SA measures. The current interface however showed almost twice as many high correlations as the ERTMS one and had the three highest correlations of all conditions, which were between saccadic duration and both reaction time and MARS score (both of which were significant), and fixation duration and SPAM outcomes. The high negative correlation between fixation duration and SPAM accuracy is supported by findings that longer fixations are representative of lower SA (Petersen et al., 2019; Miramonte, 2017), and supports the idea of SPAM and fixation duration representing SA. However, since this finding is not significant further investigation needs to be conducted. With regards to the significant correlations occurring in both interfaces in conditions 2a and 2b, eye tracking measures only showed significant correlations with reaction times and self-rated MARS scores. Specifically, between saccadic durations and both reaction times and MARS, and between MARS and aggregate saccades in the current interface. In the ERTMS interface significant associations were found between reaction time and both saccadic amplitude and fixations, as well as between aggregate fixations and saccades. These findings point to aggregate saccades and saccadic duration being promising SA measures for comparing a control and new interface after a learning trial, but also raise the issue that since SPAM didn’t significantly correlate with any of the eye tracking measures in these two non-adaptation conditions, perhaps this is not the case especially since SPAM averages were found

to significantly differ between the two interfaces in 2a and 2b. However, given that SPAM did significantly correlate with MARS ratings, it is likely that further investigation with a larger sample size would show more clear and sound associations.

3.4.3. Limitations

The current experiment was based upon a small sample size of train traffic controllers but given that the current experiment was conducted during the period of Covid 19 only a small sample size could be recruited and likewise tested during a specific time period and following specific pandemic protocol. Likewise, all preparation for the study, including quite complex tasks such as operationalizing the situations for the different conditions and constructing the SPAM assessment all had to be conducted online, rather than in person between researchers. It is also worth noting that the novel interface used in the current case study was based on what operators are used to seeing at present in their day to day workplace, and thus only functional ERTMS relevant design features were investigated in terms of visual salience. Finally the novel service area operators worked in during the simulation seemed to have more of an impact than was anticipated.

3.4.4. Summary and Future Recommendations

The use of qualitative human factors techniques alongside quantitative human factors techniques gave a more extensive insight into how ERTMS can be developed for use in the future, based on both clarified workflows as well as sound statistical conclusions. Support was shown for SA measurement techniques such as self-rating measures and probe queries. Additionally, support was shown for both saccadic and fixation eye tracking measures in representing SA when assessing control room interface design. It is recommended that the current study is replicated in the future with a larger sample size, and with more training regarding unfamiliar service areas prior to testing to make the process more efficient. This would likely have the positive effect of increasing end user (TTC) uptake and trust in the newly implemented system. Likewise, it is recommended that the array of eye tracking measures used in the current experiment are further investigated in control room contexts in association with SA specifically, and also compared with MWL to further establish which measures tend more to one construct or the other and to explore if the findings from the current experiment can be replicated or more concretely explored.

3.5. Conclusion

This thesis presents a conceptual framework for integrating quantitative human factors techniques into control room design for usability troubleshooting.

The conceptual model was developed by systematically appraising and arranging different aspects of control room systems and human factors research techniques into a framework that is suitable for guiding more human centered control room design. A subsection of the conceptual framework has been used to analyze a novel interface compared to a control one and demonstrated the plausibility of incorporating quantitative human factors research techniques more extensively to support more effective control room design processes.

The conceptual framework presented in the current thesis was intended for future use for control room design processes by human factors engineers working in safety critical domains. The application of this conceptual framework is intended to direct control room design towards a more human centric focus, and thus ensure more effective and safe control room systems design. This purpose was fulfilled through the ERTMS case study, which demonstrated the usefulness of incorporating a wider range of research techniques, by adding dimensionality to SA level outcomes for TTCs by showing trends in associations between eye-tracking metrics and SA measures when comparing different interfaces. However this dimensionality would likely be further demonstrated in a more controlled research setting with a larger sample size. It is thus recommended that these findings are replicated in future control room research and that the Conceptual Framework for Control Room Design is applied in its entirety in future research.

References

Agost, M., & Vergara, M. (2010). A CONCEPTUAL FRAMEWORK FOR IMPRESSIONS ELICITED IN HUMAN- PRODUCT INTERACTION. DESIGN. *International Conference on Kansei Enginerring and Emotional Research 2010*.

Allsop, J., Gray, R., Bülthoff, H. H., & Chuang, L. (2017). Effects of Anxiety and cognitive load on instrument scanning behavior in a flight simulation. *Proceedings of the 2nd Workshop on Eye Tracking and Visualization, ETVIS 2016*, 55–59.

Andersson, A.-P., & Cappelen, B. (1999). *Ambiguity-A User Quality Collaborative Narrative in a Multimodal User Interface*. www.aaai.org

Andersson, M and Lutzhoft, M, Engine control rooms - Human factors, *Proceedings of the 2007 RINA, Royal Institution of Naval Architects International Conference - Human Factors in Ship Design, Safety and Operation, 2007*, London, United kingdom, pp. 65-68.

Andersson, J. (2010). *THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING A Conceptual Model for Analysis of Automation Usability Problems in Control Room Settings*.

Andersson, J. (2010). A Conceptual Model for Analysis of Automation Usability Problems in Control Room Settings. *Research Series from Chalmers University of Technology, Department of Product and Production Development: Report*, 63. http://publications.lib.chalmers.se/records/fulltext/121920.pdf

Andriole, S. J. (1989). *Storyboard prototyping: a new approach to user requirements analysis*. 280.

Atkins, M. S., Jiang, X., Tien, G., & Zheng, B. (2012). Saccadic delays on targets while watching videos. *Eye Tracking Research and Applications Symposium (ETRA)*, 405–408.

Babaei Pouya, A., Hazrati, S., Vosoughi, M., Mosavianasl, Z., & Habibi, E. (2017) *Evaluation Human Error in Control Room* (Vol. 11, Issue 4).

Babaeipouya, A., Hazrati, S., Vosoughi, M., Mosavianasl, Z., & Habibi, E. (2018). An evaluation of human error in control room. *Annals of Tropical Medicine and Public Health, 1.2 Special Issue*, SP20.

Bannon, L. J., & Boedker, S. (1989). Beyond the Interface: Encountering Artifacts in Use. *Computer Science Department Aarhus*.

Bannon, L. J. (1995). From Human Factors to Human Actors: The Role of Psychology and Human-Computer Interaction Studies in System Design. *Readings in Human–Computer Interaction*, 205–214.

Barbieri, L., Angilica, A., Bruno, F., & Muzzupappa, M. (2013). Mixed prototyping with configurable physical archetype for usability evaluation of product interfaces. *Computers in Industry*, *64*(3), 310–323.

Barnum, C. (2019). *The State of UX Research* (Vol. 15).

Baysari, M. T., McIntosh, A. S., & Wilson, J. R. (2008). Understanding the human factors contribution to railway accidents and incidents in Australia. *Accident Analysis and Prevention*, *40*(5), 1750–1757. https://doi.org/10.1016/j.aap.2008.06.013

Bernhaupt, R., Palanque, P., Winckler, M., & Navarre, D. (2007). Usability study of multi-modal interfaces using eye-tracking. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4663 LNCS*(PART 2), 412–424.

Bennett, S. (1993). *A History of Control Engineering 1930-1955*. London: Peter Peregrinus Ltd. On behalf of the Institution of Electrical Engineers.

Bjørneseth, F. B., Renganayagalu, S. K., Dunlop, M. D., Hornecker, E., & Komandur, S. (2012). *Towards an Experimental Design Framework for Evaluation of Dynamic Workload and Situational Awareness in Safety Critical Maritime Settings*. *2006*. https://doi.org/10.14236/ewic/hci2012.43

Bjørneseth, F. B., Renganayagalu, S. K., Dunlop, M. D., Hornecker, E., & Komandur, S. (2012, September 1). *Towards an Experimental Design Framework for Evaluation of Dynamic Workload and Situational Awareness in Safety Critical Maritime Settings*. https://doi.org/10.14236/ewic/HCI2012.43

Blascheck, T., Kurzhals, K., Raschke, M., Strohmaier, S., Weiskopf, D., & Ertl, T. (2016). AOI hierarchies for visual exploration of fixation sequences. *Eye Tracking Research and Applications Symposium (ETRA)*, *14*, 111–118. https://doi.org/10.1145/2857491.2857524

Boring, R. L. (2007). Dynamic Human Reliability Analysis: Benefits and Challenges of Simulating Human Performance. *Proceedings of the European Safety and Reliability Conference*.

Boring, R. L. (2017). As low as reasonable assessment (ALARA): Applying discount usability to control room verification and validation. *Risk, Reliability and Safety: Innovating Theory and Practice - Proceedings of the 26th European Safety and Reliability Conference, ESREL 2016*, *September*, 153.

Boring, R. L. (2014). Human reliability analysis for digital human-machine interfaces: A wish list for future research. *PSAM 2014 - Probabilistic Safety Assessment and Management*, *July*.

Boring, R. L., & Joe, J. C. (2015). Baseline evaluations to support control room modernization at nuclear power plants. *9th International Topical Meeting on Nuclear Plant Instrumentation, Control, and Human-Machine Interface Technologies, NPIC and HMIT 2015*, *2*, 911–922.

Boring, R. L., & Gertman, D. I. (2005). Advancing Usability Evaluation Through Human Reliability Analysis. *Human Computer Interaction International 2005*.

Boring, R., Ulrich, T., Lew, R., Kovesdi, C., Rice, B., Poresky, C., Spielman, Z., & Savchenko, K. (2017). Analog, Digital, or Enhanced Human-System Interfaces? Results of an Operator-in-the-Loop Study on Main Control Room Modernization for a Nuclear Power Plant. In *Idaho National Lab. (INL), Idaho Falls, ID (United States)*.

Boy, G. A. (1998). Cognitive function analysis for human-centered automation of safety-critical systems. *Conference on Human Factors in Computing Systems - Proceedings*, 265–272.

Braseth, A. O., Nihlwing, C., Svengren, H., Veland, Ø., Hurlen, L., & Kvalem, J. (2009). Lessons learned from Halden project research on human system interfaces. *Nuclear Engineering and Technology*, *41*(3), 215–224.

Browning, T. R., Sage, A. P., & Rouse, W. B. (2009). *Handbook of Systems Engineering and Management, 2 nd Edition*.

Bruemmer, D. J., Few, D. A., Boring, R. L., Marble, J. L., Walton, M. C., & Nielsen, C. W. (2005). Shared understanding for collaborative control. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.*, *35*(4), 494–504.

Buettner, R., Sauer, S., Maier, C., & Eckhardt, A. (2018). Real-time Prediction of User Performance based on Pupillary Assessment via Eye-Tracking. *AIS Transactions on Human-Computer Interaction*, *10*(1), 26–60.

Burns, C. M., Skraaning, G., Jamieson, G. A., Lau, N., Kwok, J., Welch, R., & Andresen, G. (2008). Evaluation of ecological interface design for nuclear process control: Situation awareness effects. *Human Factors*, *50*(4), 663–679.

Burns, C. M., Skraaning, G., Jamieson, G. A., Lau, N., Kwok, J., Welch, R., & Andresen, G. (2008). Evaluation of ecological interface design for nuclear process control: Situation awareness effects. *Human Factors*, *50*(4), 663–679.

Burns, C., Jamieson, G., Skraaning, G., Lau, N., & Kwok, J. (2007). Supporting situation awareness through ecological interface design. *Proceedings of the Human Factors and Ergonomics Society*, *1*(October), 205–209.

Butscher, S., Mueller, J., Schwarz, T., & Reiterer, H. (2013). Blended Interaction as an Approach for Holistic Control Room Design. *To Appear: In Proceedings of the Workshop on Blended Interaction: Envisioning Future Collaborative Interactive Spaces (In Conjunction with CHI 2013 Conference)*.

Camburn, B., Viswanathan, V., Linsey, J., Anderson, D., Jensen, D., Crawford, R., Otto, K., & Wood, K. (2017). Design prototyping methods: State of the art in strategies, techniques, and guidelines. *Design Science*, *3*(Schrage 1993), 1–33.

Chang, E., & Dillon, T. S. (1997). *3 Automated Usability Testing*.

Chang, E., & Dillon, T. S. (1997). Automated Usability Testing. *Human-Computer Interaction INTERACT '97*, *76*(January), 77–84.

Chiappe, D., Vu, K.-P. L., Strybel, T. Z., Manke, B., & By Adriana J Miramontes BA, P. D. (2017). *EXAMINING EYE FIXATION PATTERNS DURING THE SITUATION PRESENT ASSESSMENT METHOD (SPAM) UNDER VARYING LEVELS OF WORKLOAD*.

Coyne, J., & Sibley, C. (2016). Investigating the use of two low cost Eye tracking systems for detecting pupillary response to changes in mental workload. *Proceedings of the Human Factors and Ergonomics Society*, 37–41.

Crawford, E. G., Toft, Y., & Kift, R. L. (2014). Attending to technology adoption in railway control rooms to increase functional resilience. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8532 LNAI*, 447–457.

Crawford, E. G. C., & Kift, R. L. (2018). Keeping track of railway safety and the mechanisms for risk. *Safety Science*, *110*, 195–205.

Crawford, E. G., Toft, Y., & Kift, R. L. (2013). New control room technologies: human factors analytical tools for railway safety. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, *227*(5), 529–538.

Crawford, E., Toft, Y., Kift, R., & Crawford, C. (2010). Entering the conceptual age: implications for control room operators and safety. *Applied Ergonomics*, *May 2014*.

Czaja, S.J. & Nair, S.N. (2012). Human Factors Engineering and Systems Design. In G. Salvendy, *Human Factors Handbook* (pp. 38-56). Hoboken, New Jersey: John Wiley and Sons.

Dadashi, N., Wilson, J. R., Sharples, S., Golightly, D., & Clarke, T. (2011). A framework of data processing for decision making in railway intelligent infrastructure. *2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, CogSIMA 2011*, *February*, 276–283.

Dadashi, N., Wilson, J. R., Golightly, D., & Sharples, S. (2014). A framework to support human factors of automation in railway intelligent infrastructure. *Ergonomics*, *57*(3), 387–402. https://doi.org/10.1080/00140139.2014.893026

Dalinger, I., Smurov, M., Sukhikh, N., & Tsybova, E. (2016). Pilot's situational awareness and methods of its assessment. *Indian Journal of Science and Technology*, *9*(46).

Dar, A. H., Wagner, A. S., & Hanke, M. (2021). REMoDNaV: robust eye-movement classification for dynamic stimulation. *Behavior Research Methods*, *53*(1), 399–414.

De Carvalho, P. V. R. (2006). Ergonomic field studies in a nuclear power plant control room. *Progress in Nuclear Energy*, *48*(1), 51–69.

De Felice, F., & Petrillo, A. (2011). Methodological approach for performing human reliability and error analysis in railway transportation system. *International Journal of Engineering and Technology*, *3*(5), 341–353.

De Greef, T., Lafeber, H., Van Oostendorp, H., & Lindenberg, J. (2009). Eye movement as indicators of mental workload to trigger adaptive automation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5638 LNAI*, 219–228.

Dehais, F., Lafont, A., Roy, R., & Fairclough, S. (2020). A Neuroergonomics Approach to Mental Workload, Engagement and Human Performance. *Frontiers in Neuroscience*, *14*(April), 1–17.

Desvergez, A., Winer, A., Gouyon, J. B., & Descoins, M. (2019). An observational study using eye tracking to assess resident and senior anesthetists' situation awareness and visual perception in postpartum hemorrhage high fidelity simulation. *PLoS ONE*, *14*(8).

Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., & Holmqvist, K. (2012). It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior Research Methods*, *44*(4), 1079–1100.

Di Nocera, F., Terenzi, M., & Camilli, M. (2015). Another Look at Scanpath: Distance to Nearest Neighbour As a Measue of Mental Workload. *Developments in Human Factors in Transportation, Design & Evaluation, July 2015*, 295–303.

Di Stasi, L. L., Álvarez-Valbuena, V., Cañas, J. J., Maldonado, A., Catena, A., Antolí, A., & Candido, A. (2009). Risk behaviour and mental workload: Multimodal assessment techniques applied to motorbike riding simulation. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(5), 361–370.

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

Dias, F. J. (2017). *Pattern Based Usability Testing*. FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO.

Diaz-Piedra, C., Rieiro, H., Cherino, A., Fuentes, L. J., Catena, A., & Di Stasi, L. L. (2019). The effects of flight complexity on gaze entropy: An experimental study with fighter pilots. *Applied Ergonomics*, *77*, 92–99.

Dicks, R. S. (2002). *Mis-Usability: On the Uses and Misuses of Usability Testing*. http://www.useit.com/

Dijk, H. Van. (2010). *Executive summary The Applicability of Eye Movements as an Indicator of Situation Awareness in a Flight Simulator Experiment*. 1–20.

Djamasbi, S; Tullis, M; Dai, R. (2010). Efficiency, Trust, and Visual Appeal: Usability Testing through Eye Tracking. *Hawaii International Conference on System Sciences*, 1–10.

dos Santos, I. J. A. L., Teixeira, D. V., Ferraz, F. T., & Carvalho, P. V. R. (2008). The use of a simulator to include human factors issues in the interface design of a nuclear power plant control room. *Journal of Loss Prevention in the Process Industries*, *21*(3), 227–238.

Drusch, G., & Bastien, J. M. C. (2012). Analyzing visual scanpaths on the Web using the mean shift procedure and T-pattern detection: A bottom-up approach. *ACM International Conference Proceeding Series*, 181–184.

Dubois, L., Forêt, J. M., Mack, P., & Ryckaert, L. (2010). Advanced logic for alarm and event processing: Methods to reduce cognitive load for control room operators. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, *11*(PART 1).

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., & Danko, N. (1999). *Situation Awareness As a Route Air Traffic Controllers*. 15.

Durso, F. T., & Sethumadhavan, A. (2008). Situation awareness: Understanding dynamic environments. In *Human Factors* (Vol. 50, Issue 3, pp. 442–448).

Durso, F. T., Kathtyn Bleckley, M., & Dattel, A. R. (2006). *Does Situation Awareness Add to the Validity of Cognitive Tests?*

Durso, F.T. and Alexander, A.L.(2010). *Managing Workload, Performance and Situation Awareness in Aviation Systems.* Human Factors in Aviation. 2. 217-247.

Ehmke, C., & Wilson, S. (2007). *Identifying Web Usability Problems from Eyetracking Data*.

Elliott, A. C, "Human Factors for railway signalling and control systems," *IET Professional Development Course on Railway Signalling and Control Systems (RSCS 2010)*, 2010, pp. 237-249

Endsley, M. R., Sollenberger, R., & Stein, E. (2000). Situation awareness: A comparison of measures. *Human Performance, Situation Awareness and Automation: User-Centered Design for the New Millennium, November*.

Endsley, M.R. (2012). Human Factors Engineering and Systems Design. In G. Salvendy, *Human Factors Handbook* (pp.553-568). Hoboken, New Jersey: John Wiley and Sons. Endsley, M. R. (2018). Automation and situation awareness. *Automation and Human Performance: Theory and Applications*, 163–181.

Endsley, M. R. (2015). Situation awareness misconceptions and misunderstandings. *Journal of Cognitive Engineering and Decision Making*, *9*(1), 4–32.

Endsley, M. R., & Rodgers, M. D. (1996). *ATTENTION DISTRIBUTION AND SITUATION AWARENESS IN AIR TRAFFIC CONTROL*.

Fabio, R. A., Incorpora, C., Errante, A., Mohammadhasni, N., Capri, T., Carrozza, C., Sandro, D. S., & Falzone, A. (2015). The Influence of Cognitive Load and Amount of Stimuli on Entropy through Eye tracking measures. *EuroAsianPacific Joint Conference on Cognitive Science*, *2010*, 199–204.

Fabo, P; Durikovic, R. (2012). Automated usability measurement of arbitrary desktop application with eyetracking. *International Conference on Information Visualisation*.

Falkland, E. C., & Wiggins, M. W. (2019). Cross-task cue utilisation and situational awareness in simulated air traffic control. *Applied Ergonomics*, *74*(July 2018), 24–30.

Fallahi, M., Motamedzade, M., Heidarimoghadam, R., Soltanian, A. R., & Miyake, S. (2016). Effects of mental workload on physiological and subjective responses during traffic density monitoring: A field study. *Applied Ergonomics*, *52*, 95–103.

Feuerstack, S., Blumendorf, M., Kern, M., Kruppa, M., Quade, M., Runge, M., & Albayrak, S. (2008). Automated usability evaluation during model-based interactive system development. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5247 LNCS*, 134–141.

Foy, H. J., & Chapman, P. (2018). Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. *Applied Ergonomics*, *73*(June 2017), 90–99.

Foyle, D. C., Goodman, A., & Hooey, B. L. (2003). *NASA Aviation Safety Program Conference on Human Performance Modeling of Approach and Landing with Augmented Displays*.

Fraser, J., & Plewes, S. (2015). Applications of a UX Maturity Model to Influencing HF Best Practices in Technology Centric Companies – Lessons from Edison. *Procedia Manufacturing*, *3*, 626–631.

Froehlich, J., Findlater, L., Landay, J., & Science, C. (2010). *The Design of Eco-Feedback Technology*. 1999–2008.

Fujino M, Lee J, Hirano T, Saito Y, Itoh M. Comparison of SAGAT and SPAM for Seeking Effective Way to Evaluate Situation Awareness and Workload During Air Traffic Control Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 2020;64(1):1836-1840.

Furniss, D., Curzon, P., & Blandford, A. (2018). Exploring organisational competences in Human Factors and UX project work: managing careers, project tactics and organisational strategy. *Ergonomics*, *61*(6), 739–761.

Gartenberg, D., McCurry, M., & Trafton, G. (2011). Situation awareness reacquisition in a supervisory control task. *Proceedings of the Human Factors and Ergonomics Society, September 2011*, 355–359.

Gaver, W. W., Beaver, J., & Benford, S. (2003). Ambiguity as a resource for design. *Conference on Human Factors in Computing Systems - Proceedings*, *5*, 233–240.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise Differences in the Comprehension of Visualizations: A Meta-Analysis of Eye-Tracking Research in Professional Domains. *Educational Psychology Review*, *23*(4), 523–552.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise Differences in the Comprehension of Visualizations: a Meta-Analysis of Eye-Tracking Research in Professional Domains. *Educational Psychology Review*, *23*(4), 523–552.

Gerber, M. A., Schroeter, R., & Vehns, J. (2019). A video-based automated driving simulator for automotive UI prototyping, UX and behaviour research. *Proceedings - 11th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2019*, 14–23.

Goh, J. T., Hu, S., & Fang, Y. (2019). Human-in-the-loop simulation for crane lift planning in modular construction on-site assembly. In C. Wang, Y. K. Cho, F. Leite, & A. Behzadan (Eds.), *Computing in Civil Engineering 2019: Visualization, Information Modeling, and Simulation - Selected Papers from the ASCE International Conference on Computing in Civil Engineering 2019* (pp. 71-78). (Computing in Civil Engineering 2019: Visualization, Information Modeling, and Simulation - Selected Papers from the ASCE International Conference on Computing in Civil Engineering 2019). American Society of Civil Engineers.

Goldberg, J. H., & Helfman, J. I. (2010). Scanpath clustering and aggregation. *Eye Tracking Research and Applications Symposium (ETRA)*, 227–234.

Goldberg, J. H., & Wichansky, A. M. (2003). *Eye tracking in usability evaluation: A practitioner's guide Optimal Product Bundling in Recommender Systems View project Eye Tracking View project*. https://www.researchgate.net/publication/259703518

Gray, C. M., & Chivukula, S. S. (2019, May 2). Ethical mediation in UX practice. *Conference on Human Factors in Computing Systems - Proceedings*.

Greenberg, S. ;, Buxton, B., & Greenberg, S. (2007). *Usability Evaluation Considered Harmful(Some of the Time) Usability Evaluation Considered Harmful (Some of the Time)*. http://hdl.handle.net/1880/45915unknownDownloadedfromPRISM:https://prism.ucalgary.ca

Grossman, T., Fitzmaurice, G., & Attar, R. (2009). A survey of software learnability: Metrics, methodologies and guidelines. *Conference on Human Factors in Computing Systems - Proceedings*, 649–658.

Gu, Z., Jin, C., Chang, D., & Zhang, L. (2020). Predicting webpage aesthetics with heatmap entropy. *Behaviour and Information Technology*, 1–22.

Ha, C. H., Kim, J. H., Lee, S. J., & Seong, P. H. (2006). Investigation on relationship between information flow rate and mental workload of accident diagnosis tasks in NPPs. *IEEE Transactions on Nuclear Science*, *53*(3), 1450–1459.

Hall, N., Lowe, C., & Hirsch, R. (2015). Human Factors Considerations for the Application of Augmented Reality in an Operational Railway Environment. *Procedia Manufacturing*, *3*, 799–806.

Hareide, O. S., & Ostnes, R. (2017). Scan Pattern for the Maritime Navigator. *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation*, *11*(1), 39–47.

Hareide, O. S., & Ostnes, R. (2018). Validation of a maritime usability study with eye tracking data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10916 LNAI*(2), 273–292.

Hareide, O., & Ostnes, R. (2017). *Maritime usability study by analysing Eye Tracking data*. *70*(5), 927–943.

Hareide, O., Ostnes, R., & Mjelde, F. (2016). Understanding the Eye of the Navigator. *Journal on Marine Navigation and Safety*, *11*(1).

Hartson, H. R. (2003). Cognitive, physical, sensory, and functional affordances in interaction design. *Behaviour and Information Technology*, *22*(5), 315–338.

Hartson, H. R. (2003). Cognitive, physical, sensory, and functional affordances in interaction design. *Behaviour and Information Technology*, *22*(5), 315–338.

Hartson, H. R. (2003). Cognitive, physical, sensory, and functional affordances in interaction design. *Behaviour and Information Technology*, *22*(5), 315–338.

Hasanzadeh, S., Esmaeili, B., & Dodd, M. D. (2016). Measuring Construction Workers' Real-Time Situation Awareness Using Mobile Eye-Tracking. *Construction Research Congress 2016*, 2894–2904.

Hasanzadeh, S., Esmaeili, B., & Dodd, M. D. (2018). Examining the Relationship between Construction Workers' Visual Attention and Situation Awareness under Fall and Tripping Hazard Conditions: Using Mobile Eye Tracking. *Journal of Construction Engineering and Management*, *144*(7).

Hasanzadeh, S., Esmaeili, B., & Dodd, M. D. (2017). Measuring the Impacts of Safety Knowledge on Construction Workers' Attentional Allocation and Hazard Detection Using Remote Eye-Tracking Technology. *Journal of Management in Engineering*, *33*(5), 04017024.

Hendrie, M., Alvarez, I., & Hooker, B. (2015). Prototyping adaptive automotive UX : a Design Pedagogy approach. *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '15*.

Hessels, R. S., Kemner, C., van den Boomen, C., & Hooge, I. T. C. (2016). The area-of-interest problem in eyetracking research: A noise-robust solution for face and sparse stimuli. *Behavior Research Methods*, *48*(4), 1694–1712

Hidalgo, J., Genaidy, A., Karwowski, W., Christensen, D., Huston, R., & Stambough, J. (1997). A comprehensive lifting model: Beyond the NIOSH lifting equation. *Ergonomics*, *40*(9), 916–927.

Hollnagel, E (2012). Human Factors Engineering and Systems Design. In G. Salvendy, *Human Factors Handbook* (pp. 383-396). Hoboken, New Jersey: John Wiley and Sons.Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Halszka, J., & van de Weijer, J. (2011). *Eye Tracking : A Comprehensive Guide to Methods and Measures*. Oxford University Press.

Hong, T. C., Andrew, H. S. Y., & Kenny, C. W. L. (2015). Assessing the situation awareness of operators using maritime augmented reality system (MARS). *Proceedings of the Human Factors and Ergonomics Society*, *2015-January*, 1722–1726.

Hooge, I., & Camps, G. (2013). Scan path entropy and arrow plots: Capturing scanning behavior of multiple observers. *Frontiers in Psychology*, *4*(DEC), 1–10.

Horchers, J. O. (2001). A pattern approach to interaction design. *AI and Society*, *15*(4), 359–376.

Hornbaek, K., & Stage, J. (2006). *The Interplay Between Usability Evaluation and User Interaction Design*.

Hornbæk, K., & Stage, J. (2006). The interplay between usability evaluation and user interaction design. *International Journal of Human-Computer Interaction*, *21*(2), 117–123.

Hornbæk, K., & Stage, J. (2006). The interplay between usability evaluation and user interaction design. *International Journal of Human-Computer Interaction*, *21*(2), 117–123.

Hosseini-khayat, A., & Canada, A. B. (n.d.). *Low-Fidelity Prototyping of Gesture-based Applications*. 289–294.

Hosseini-Khayat, A., Hellmann, T. D., & Maurer, F. (2010). Distributed and Automated Usability Testing of Low-Fidelity Prototypes. *2010 Agile Conference*, 59–66.

Hosseini-Khayat, A., Hellmann, T. D., & Maurer, F. (n.d.). *Distributed and Automated Usability Testing of Low-Fidelity Prototypes*.

Hu, W. L., Rivetta, C., MacDonald, E., & Chassin, D. P. (2019). Modeling of Operator Performance for Human-in-the-loop Power Systems. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11571 LNAI*, 39–54.

Hyönä, J. (Jukka), Radach, R. (Ralph), & Deubel, H. (2003). Eye Tracking in Human–Computer Interaction and Usability Research: Ready to Deliver the Promises. *The Mind's Eye : Cognitive and Applied Aspects of Eye Movement Research*, 741.

Ikuma, L. H., Harvey, C., Taylor, C. F., & Handal, C. (2014). A guide for assessing control room operator performance using speed and accuracy, perceived workload, situation awareness, and eye tracking. *Journal of Loss Prevention in the Process Industries*, *32*(December), 454–465.

Ingram, A., Wang, X., & Ribarsky, W. (2012). Towards the establishment of a framework for intuitive multi-touch interaction design. *Proceedings of the Workshop on Advanced Visual Interfaces AVI*, 66–73.

Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., & Bailey, B. P. (2005). Towards an index of opportunity: Understanding in mental workload during task execution. *CHI 2005: Technology, Safety, Community: Conference Proceedings - Conference on Human Factors in Computing Systems*, 311–320.

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). Task-evoked pupillary response to mental workload in human-computer interaction. *Conference on Human Factors in Computing Systems - Proceedings*, 1477–1480.

Itoh, K., Arimoto, M., & Akachi, Y. (2001). Eye-tracking applications to design of a new train interface for the Japanese high-speed railway. In *Usability evaluation and interface …* (p. 5). http://www.me.titech.ac.jp/~itohlab/ronbun_copy/c2-2001-itoh.pdf

Jaewhan Kim Seung-Cheol Jang, and Jong-Bae Wang, W. J. (2006). A Case Study for the Selection of a Railway Human Reliability Analysis Method. *International Railway Safety Conference*, *January*. http://www.intlrailsafety.com/Dublin/Additional_Papers/Kim_JW.pdf

Jannat, M., Hurwitz, D. S., Monsere, C., & Funk, K. H. (2018). The role of driver's situational awareness on right-hook bicycle-motor vehicle crashes. *Safety Science*, *110*(September 2017), 92–101.

Jansen, R. J., Sawyer, B. D., Van Egmond, R., De Ridder, H., & Hancock, P. A. (2016). Hysteresis in Mental Workload and Task Performance. *Human Factors*, *58*(8), 1143–1157.

Jimenez-Molina, A., Retamal, C., & Lira, H. (2018). Using psychophysiological sensors to assess mental workload during web browsing. *Sensors (Switzerland)*, *18*(2), 1–26.

Johnsen, S. O., & Veen, M. (2013). Risk Assessment of Critical Communication Infrastructure in Railways in Norway. *Cognition, Technology and Work*.

Jou, Y.-T., Yenn, T.-C., Lin, C. J., Yang, C.-W., & Chiang, C.-C. (2009). Evaluation of operators' mental workload of human–system interface automation in the advanced nuclear power plants. *Nuclear Engineering and Design*, *239*(11), 2537–2542.

Kantowitz, B. H. (1992). Selecting measures for human factors research. *Human Factors*, *34*(4), 387–398.

Kantowitz BH. (2000) Attention and Mental Workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 2000;44(21):3-456-3-459.

Karltun, A., & Karltun, J. (2019). Benefits of the Human-Technology-Organization Concept in Teaching Ergonomics – Students Perspective. *Advances in Intelligent Systems and Computing*, *821*, 627–636.

Karltun, A., & Karltun, J. (2019). *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (eds.); Vol. 821, Issue August, pp. 627–636). Springer International Publishing.

Karwowski, W. (2012). Human Factors Engineering and Systems Design. In G. Salvendy, *Human Factors Handbook* (pp.4-36). Hoboken, New Jersey: John Wiley and Sons.Katrahmani, A., Ahmadi, N., & Romoser, M. (2017). Using Situation Awareness as a Measure of Driver Hazard Perception Ability. *2017 Driving Assessment Conference*, 256–262.

Kauppi, A., Wikström, J., Sandblad, B., & Andersson, A. W. (2006). Future train traffic control: Control by re-planning. *Cognition, Technology and Work*, *8*(1), 50–56.

Kauppi, A., Wikström, J., Sandblad, B., & Andersson, A. W. (2006). Future train traffic control: Control by re-planning. *Cognition, Technology and Work*, *8*(1), 50–56.

Kim, J. H., Gunn, D. V., Schuh, E., Phillips, B. C., Pagulayan, R. J., & Wixon, D. (2008). Tracking Real-Time User Experience (TRUE): A comprehensive instrumentation solution for complex systems. *Conference on Human Factors in Computing Systems - Proceedings*, 443–451.

Kim, S. K., Suh, S. M., Jang, G. S., Hong, S. K., & Park, J. C. (2012). Empirical research on an ecological interface design for improving situation awareness of operators in an advanced control room. *Nuclear Engineering and Design*, *253*, 226–237.

Kitamura, M., Fujita, Y., & Yoshikawa, H. (2005). Review of international standards related to the design for control rooms on nuclear power plants. *Journal of Nuclear Science and Technology*, *42*(4), 406–417.

Kjeldskov, J., & Skov, M. B. (2003). Creating Realistic Laboratory Settings: Comparative Studies of Three Think-Aloud Usability Evaluations of a Mobile System. *Computer*, *June 2014*, 663–670. http://www.cs.aau.dk/~jesper/pdf/papers/Interact03-comparativeEvaluations-final.pdf

Klemmer, S. R., Hartmann, B., & Takayama, L. (2006). How bodies matter: Five themes for interaction design. *Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, DIS*, *2006*, 140–149.

Kluge, A., Salman N., & Manca, D. (2014) *Advanced Applications in Process Control and Training Needs of Field and Control Room Operators*, IIE Transactions on Occupational Ergonomics and Human Factors, 2:3-4, 121-136

Kluth, W., Krempels, K. H., & Samsel, C. (2014). Automated usability testing for mobile applications. *WEBIST 2014 - Proceedings of the 10th International Conference on Web Information Systems and Technologies*, *2*, 149–156.

Kornaszewski, M. (2018). *MICROPROCESSOR TECHNOLOGY AND PROGRAMMABLE LOGIC CONTROLLERS IN NEW GENERATION RAILWAY TRAFFIC CONTROL AND MANAGEMENT SYSTEMS Telematics Telematics Transport System Transport System Microprocessor Technology and Programmable Logic Controllers in New Generation Railway Traffic Control and Management Systems*.

Kovesdi, C., Spielman, Z., LeBlanc, K., Rice, B. (2017). APPLICATION OF EYE TRACKING FOR MEASUREMENT AND EVALUATION IN HUMAN FACTORS STUDIES IN CONTROL ROOM MODERNIZATION Kovesdi,. *Control and Human Machine Interface*, *10*.

Kovesdi, C., Spielman, Z., LeBlanc, K., & Rice, B. (2018). Application of Eye Tracking for Measurement and Evaluation in Human Factors Studies in Control Room Modernization. *Nuclear Technology*, *202*(2–3), 220–229.

Kramer R, Johnson A, Zeilstra MP. The Integrated Workload Scale – Translation and validation of a subjective workload scale. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*. 2017;231(10):1123-1129

Krejtz, K., Duchowski, A., Szmidt, T., Krejtz, I., Perilli, F. G., Pires, A., Vilaro, A., & Villalobos, N. (2015). Gaze transition entropy. *ACM Transactions on Applied Perception*, *13*(1).

Krueger, A. (2013). A SYSTEMS APPROACH TO THE ASSESSMENT OF MENTAL WORKLOAD IN A SAFETY-CRITICAL ENVIRONMENT by. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699.

Kuutti, K., Battarbee, K., Säde, S., Mattelmäki, T., Keinonen, T., Teirikko, T., & Tornberg, A. M. (2001). Virtual prototypes in usability testing. *Proceedings of the Hawaii International Conference on System Sciences*, *January*, 134.

Kwekkeboom, E.J. (2012). *Consolidation of Refinery Control Rooms.* Petroleum Technology Quarterly. 17. 85-91. 1.

Kwee-Meier, S. T., Wiessmann, M., & Mertens, A. (2017). Integrated information visualization and usability of user interfaces for safety-critical contexts. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10276 LNAI*, 71–85.

Lee, J. D., & Sanquist, T. F. (2000). Augmenting the operator function model with cognitive operations: assessing the cognitive demands of technological innovation in ship navigation. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.*, *30*(3), 273–285.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lewis, J.R. (2012). Human Factors Engineering and Systems Design. In G. Salvendy, *Human Factors Handbook* (pp.1267-1312). Hoboken, New Jersey: John Wiley and Sons.

Li, W. C., Chiu, F. C., & Wu, K. J. (2012). *The Evaluation of Pilots Performance and Mental Workload by Eye Movement
*. *September*, 24–28.

Li, W. C., Horn, A., Sun, Z., Zhang, J., & Braithwaite, G. (2020). Augmented visualization cues on primary flight display facilitating pilot's monitoring performance. *International Journal of Human Computer Studies*, *135*.

Li, X., Powell, M. S., & Horberry, T. (2012). Human Factors in Control Room Operations in Mineral Processing: Elevating Control From Reactive to Proactive. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 88–111.

Lischke, L., Mayer, S., Preikschat, A., Schweizer, M., Vu, B., Wozniak, P. W., & Henze, N. (2018). Understanding large display environments: Contextual inquiry in a control room. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*.

Lisińska-Kuśnierz, M., & Krupa, M. (2020). Suitability of eye tracking in assessing the visual perception of architecture-A case study concerning selected projects located in cologne. *Buildings*, *10*(2).

Lo, J. (2020). *Gaming Simulation and Human Factors in Complex Socio-Technical Systems: A Multi-Level Approach to Mental Models and Situation Awareness in Railway Traffic Control*. Technische Universiteit Delft.

Lo, J. C., Sehic, E., Brookhuis, K. A., & Meijer, S. A. (2016). Explicit or implicit situation awareness? Measuring the situation awareness of train traffic controllers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *43*, 325–338.

Lounis, C., Peysakhovich, V., & Causee, M. (2020). *Lempel-Ziv Complexity of dwell sequences: visual scanning pattern differences between novice and expert aircraft pilots Christophe*.

Lowe, C. (2008). Improvements in System Safety. *Improvements in System Safety*, *October*.

Lu, Z., Coster, X., & De Winter, J. (2017). How much time do drivers need to obtain situation awareness? A laboratory-based study of automated driving. *Applied Ergonomics*, *60*, 293–304.

Maguire, M. (2013). Using human factors standards to support user experience and agile design. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8009 LNCS*(PART 1), 185–194.

Manhartsberger, M., & Zellhofer, N. (2005). *Eye tracking in usability research: What users really see* (Vol. 198).

Marble, J. L., Bruemmer, D. J., & Few, D. A. (2003). Lessons learned from usability tests with a collaborative cognitive workspace for human-robot teams. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, *1*(June 2014), 448–453.

Mark, J., Curtin, A., Kraft, A., Sands, T., Casebeer, W. D., Ziegler, M., & Ayaz, H. (2020). Eye tracking-based workload and performance assessment for skill acquisition. In *Advances in Intelligent Systems and Computing* (Vol. 953). Springer International Publishing.

Matthews, M. D., & Beal, S. A. (2002). *Assessing Situation Awareness in Field Training Exercises*.

Matthey, T., Cickovski, T., Hampton, S., Ko, A., Ma, Q., Nyerges, M., Raeder, T., Slabach, T., & Izaguirre, J. A. (2004). ProtoMol, an object-oriented framework for prototyping novel

algorithms for molecular dynamics. *ACM Transactions on Mathematical Software*, *30*(3), 237–265. https://doi.org/10.1145/1024074.1024075

Mazoni, M., & Hassenzahl, M. (n.d.). *Quantitative Qualitative Evaluation Development Work based Leisure based Personal Social Reductive Holistic Towa... Related papers Towards t he evaluat ion of UX Carmelo Ardit o T he hedonic/pragmat ic model of user experience*. http://www.cost294.org

McDonald, S., Edwards, H. M., & Zhao, T. (2012). Exploring think-alouds in usability testing: An international survey. *IEEE Transactions on Professional Communication*, *55*(1), 2–19.

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

McIntire, L. K., McKinley, R. A., Goodyear, C., & McIntire, J. P. (2014). Detection of vigilance performance using eye blinks. *Applied Ergonomics*, *45*(2 PB), 354–362.

Mentler, T., Rasim, T., Müßiggang, M., & Herczeg, M. (2018). Ensuring usability of future smart energy control room systems. *Energy Informatics*, *1*(S1).

Militello, L. G., & Hutton, R. J. B. (1998). Applied cognitive task analysis (ACTA): a practitioner's toolkit for understanding cognitive task demands. *Ergonomics*, *41*(11), 1618–1641.

Milnes, J., Ayed, B., Dhalla, F., Fishpool, G., Hill, J., Katramados, I., Martin, R., Naylor, G., O'gorman, T., Scannell, R., & Upgrade, M. (2015). MAST Upgrade-Construction Status. *Fusion Engineering and Design*, *96–97*, 42–47.

Miramonte, A. (2017). *EXAMINING EYE FIXATION PATTERNS DURING THE SITUATION PRESENT ASSESSMENT METHOD (SPAM) UNDER VARYING LEVELS OF WORKLOAD Presented to the Department of Psychology California State University , Long Beach In Partial Fulfillment of the Requirements for the* (Issue January). California State University, Long Beach In.

Moore, K., & Gugerty, L. (2010). Development of a novel measure of situation awareness: The case for eye movement analysis. *Proceedings of the Human Factors and Ergonomics Society*, *3*, 1650–1654.

Nachtwei, J. (2016). *manuscript: SAGAT vs. SPAM? It depends! An Aid for Situation Awareness Method Selection (aSAms) p. 1 of 37 SAGAT versus SPAM? It depends! An Aid for Situation Awareness Method Selection (aSAms).*

Nachtwei, J. (2016). *Aid for Situation Awareness Method Selection (aSAms) SAGAT versus SPAM ? It depends ! An Aid for Situation Awareness Method Selection (aSAms) Author*

Jens Nachtwei Affiliation Corresponding Author ' s Contact Information Engineering Psychology / Cognit. *August*.

Neerincx, M. A., Kennedie, S., Grootjen, M., & Grootjen, F. (2009). Modeling the cognitive task load and performance of naval operators. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5638 LNAI*(July), 260–269.

Nijland, F., Gkiotsalitis, K., & van Berkum, E. C. (2021). Improving railway maintenance schedules by considering hindrance and capacity constraints. *Transportation Research Part C: Emerging Technologies*, *126*

Nilsson, J., & Bergman, F. (2006). *Usability Study of Interactive Decision Support and Touch Screen Technology in Control Room Environments*. 112.

Nissinen, T. (2015). *User experience prototyping – a literature review*. 1–21.

Noah, B., & Rothrock, L. (2015). *Using eye tracking for live measures of workload in a refinery control room process monitoring task.* https://www.asmconsortium.net/Documents/NoahandRothrock_2015_Using eye tracking for live measures of workload in a refinery control room process monitoring task.pdf

Norman, K. L., & Panizzi, E. (2006). Levels of automation and user participation in usability testing. *Interacting with Computers*, *18*(2), 246–264.

Okazaki, T., Murai, K., Mitomo, N., & Hikida, K. (2007). A study on navigator's mental workload in ship handling simulator. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, July*, 3644–3649.

Othman, N., & Romli, F. I. (2016). Mental workload evaluation of pilots using pupil dilation. *International Review of Aerospace Engineering*, *9*(3), 80–84.

Oxstrand, J., & Le Blanc, K. L. (2014). *Effects of Levels of Automation for Advanced Small Modular Reactors: Impacts on Performance, Workload, and Situation Awareness*. http://www.inl.gov

Palanque, P., Bastide, R., & Paternò, F. (1997). Formal Specification as a Tool for Objective Assessment of Safety-Critical Interactive Systems. *Human-Computer Interaction INTERACT '97*, 323–330.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making*, *2*(2), 140–160.

Peres, S. C., Pham, T., & Phillips, R. (2013). Validation of the system usability scale (sus): Sus in the wild. *Proceedings of the Human Factors and Ergonomics Society*, 192–196.

Petersen, L., Robert, L., Yang, J., & Tilbury, D. (2019). Situational Awareness, Driver's Trust in Automated Driving Systems and Secondary Task Performance. *SSRN Electronic Journal*, 1–26.

Petersen, L., Robert, L., Yang, X. J., & Tilbury, D. M. (n.d.). Situational Awareness, Driver's Trust in Automated Driving Systems and Secondary Task Performance. In *SAE International Journal of Connected and Autonomous Vehicles*.

Pfeuffer, K., Alexander, J., & Gellersen, H. (2016). Partially-indirect Bimanual input with gaze, pen, and touch for pan, zoom, and ink interaction. *Conference on Human Factors in Computing Systems - Proceedings*, 2845–2856.

Pickup, L., Wilson, J. R., Sharpies, S., Norris, B., Clarke, T., & Young, M. S. (2005). Fundamental examination of mental workload in the rail industry. *Theoretical Issues in Ergonomics Science*, *6*(6), 463–482.

Pierce, R. S., Vu, K.-P. L., Nguyen, J., & Strybel, T. Z. (2008). *The Relationship Between SPAM, Workload, and Task Performance on a Simulated ATC Task*.

Pikaar, R. N., Human, E., & Engineering, F. (2018). *Control room design and systems ergonomics*. *January 1992*.

Pilo De La Fuente, E., Mazumder, S. K., & Franco, I. G. (2014). Railway Electrical Smart Grids: An introduction to next-generation railway power systems and their operation. *IEEE Electrification Magazine*, *2*(3), 49–55.

Piquado, Tepring, et al, Ramos, L., & Specia, L. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, *47*(3), 560–569.

Ponsa, P., Vilanova, R., & Amante, B. (2009). Towards integral human-machine system conception: From automation design to usability concerns. *2009 2nd Conference on Human System Interactions*, 427–433.

Proctor, R. W., & Vu, K. P. L. (2010). Cumulative knowledge and progress in human factors. *Annual Review of Psychology*, *61*, 623–651.

Rajanen, D., Clemmensen, T., Iivari, N., Inal, Y., Rızvanoğlu, K., Sivaji, A., & Roche, A. (2017). *UX Professionals' Definitions of Usability and UX-A Comparison Between Turkey, Finland, Denmark, France and Malaysia*. 10.

Rajanen, D., Clemmensen, T., Iivari, N., Inal, Y., Rızvanoğlu, K., Sivaji, A., & Roche, A. (2017). UX professionals' definitions of usability and UX – A comparison between Turkey, Finland, Denmark, France and Malaysia. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10516 LNCS*(February), 218–239.

Rasmussen, M., & Laumann, K. (2014). The suitability of the SPAR-H "Ergonomics/HMI" PSF in a computerized control room in the petroleum industry. *PSAM 2014 - Probabilistic Safety Assessment and Management*, *June*.

Recarte, M. A., & Nunes, L. M. (2003). Mental Workload While Driving: Effects on Visual Search, Discrimination, and Decision Making. *Journal of Experimental Psychology: Applied*, *9*(2), 119–137.

Recarte, M. Á., Ṕerez, E., Conchillo, Á., & Nunes, L. M. (2008). Mental workload and visual impairment: Differences between pupil, blink, and subjective rating. *Spanish Journal of Psychology*, *11*(2), 374–385.

Reiman, T., & Oedewald, P. (2009). *Evaluating safety-critical organizations-emphasis on the nuclear industry Research 2009:12 Title: Evaluating safety-critical organizations-emphasis on the nuclear industry*. www.stralsakerhetsmyndigheten.se

Reinach, S., & Viale, A. (2006). Application of a human error framework to conduct train accident/incident investigations. *Accident Analysis and Prevention*, *38*(2), 396–406.

Rgen Sauer, J., Seibel, K., & Rü Ttinger, B. (2010). The influence of user expertise and prototype fidelity in usability tests. *Applied Ergonomics*, *41*, 130–140.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods. *Applied Psychology*, *53*(1), 61–86.

Rudd,J. Stern,K. & Isensee, S. 1996. *Low vs. high-fidelity prototyping debate.* Interactions 3, 1 (Jan. 1996), 76–85

Russell, M. C. (2005). *Hotspots and Hyperlinks: Using Eye-tracking to Supplement Usability Testing*. *7*(2). http://psychology.wichita.edu/newsurl/usabilitynews/72/eyetracking.asp

Salmon, P. M., Stanton, N. A., Walker, G. H., Jenkins, D., Ladva, D., Rafferty, L., & Young, M. (2009). Measuring Situation Awareness in complex systems: Comparison of measures study. *International Journal of Industrial Ergonomics*, *39*(3), 490–500.

Salmon, P., Stanton, N., Walker, G., & Green, D. (2006). Situation awareness measurement: A review of applicability for C4i environments. *Applied Ergonomics*, *37*(2), 225–238.

Shackel, B. (2009) *Interacting with Computers*, Usability – Context, framework, definition, design and evaluation,21, 339–346, 6.

Schiessl, M., Duda, S., Thölke, A., & Fischer, R. (2003). Eye tracking and its application in usability and media research. *MMI Interaktiv Journal*, *1439*.

Schiessl, M., Duda, S., Thölke, A., & Fischer, R. (2003). *Eye tracking and its application in usability and media research*.

Schipper, D., & Gerrits, L. (2018). Differences and similarities in European railway disruption management practices. *Journal of Rail Transport Planning and Management*, *8*(1), 42–55.

Schipper, D., & Gerrits, L. (2018). Differences and similarities in European railway disruption management practices. *Journal of Rail Transport Planning and Management*, *8*(1), 42–55.

Schwalm, M., Keinath, A., & Zimmer, H. D. (2008). Pupillometry as a method for measuring mental workload within a simulated driving task. *Human Factors for Assistance and Automation, May 2014*, 1–13. https://www.researchgate.net/profile/Maximilian_Schwalm/publication/262357754_Pupillometry_as_a_method_for_measuring_mental_workload_within_a_simulated_driving_task/links/0c960537600688a792000000/Pupillometry-as-a-method-for-measuring-mental-workload-with

Sears, A. (2002). The Human-Computer Interaction Handbook. In *The Human-Computer Interaction Handbook*.

Sellner, B. P., Hiatt, L. M., Simmons, R., & Singh, S. (2006). Attaining situational awareness for sliding autonomy. *HRI 2006: Proceedings of the 2006 ACM Conference on Human-Robot Interaction, 2006*, 80–87.

Shackel, B. (2009). Usability - Context, framework, definition, design and evaluation. *Interacting with Computers*, *21*(5–6), 339–346.

Shappell, S., Detwiler, C., Holcomb, K., Hackworth, C., Boquet, A., & Wiegmann, D. A. (2007). Human error and commercial aviation accidents: An analysis using the human factors analysis and classification system. *Human Factors*, *49*(2), 227–242.

Shneiderman, B. (2012). Handbook of Human Factors and Ergonomics (4th ed.). In *International Journal of Human-Computer Interaction* (Vol. 28, Issue 12, pp. 838–838).

Shneiderman, B. (2003). Promoting universal usability with multi-layer interface design. *Proceedings of the 2003 Conference on Universal Usability, CCU 2003*, 1–8.

Singh, R., Chandra, S., Dhusia, K., & Sharma, G. (2016). Capacitating surveillance and situational awareness with measure of visual engagement using eyetracker. *2016 International Conference on Computing, Communication and Automation (ICCCA)*, 1161–1165.

Smith, P., & Twizell, E. H. (1980). A finite element model of temperature distribution in the human torso. In *Applied Mathematical Modelling* (Vol. 4, Issue 3).

Smith, P., Blandford, A., & Back, J. (2009). Questioning, exploring, narrating and playing in the control room to maintain system safety. *Cognition, Technology and Work, 11*(4), 279–291.

Soliday, T. (2009). *Light Water Reactor Sustainability Program A Research Framework for Demonstrating Benefits of Advanced Control Room Technologies*. *December*.

Stanton, N. a. (2006). "Best Known Task Analysis Technique." *Applied Ergonomics*, *1*(3), 1–56.

Stapel, J., Hassnaoui, E., & Happee, R. (2020). Measuring Driver Perception: Combining Eye-Tracking and Automated Road Scene Perception. In *HUMAN FACTORS* (Vol. 00, Issue 0).

Stephens, A. N., Young, K. L., Logan, D. B., & Lenne, M. G. (2015). *An On-Road Study of the Impact of Roadside Advertising on Driving Performance and Situation Awareness.* Monash University

Strandvall, T. (2009). LNCS 5727 - Eye Tracking in Human-Computer Interaction and Usability Research. In *LNCS* (Vol. 5727).

Strybel, T. Z., Vu, K. P. L., Kraft, J., & Minakata, K. (2008). Assessing the situation awareness of pilots engaged in self spacing. *Proceedings of the Human Factors and Ergonomics Society*, *1*(February), 11–15.

Stuckey, H. (2013). Three types of interviews: Qualitative research methods in social health. *Journal of Social Health and Diabetes*, *01*(02), 056–059

Sturm, J., Bakx, I., Cranen, B., Terken, J., & Fusi, W. (2020). *PDF hosted at the Radboud Repository of the Radboud University Nijmegen Usability Evaluation of a Dutch Multimodal System for Train Timetable Information*.

Sturre, L., Chiappe, D., Vu, K. P. L., & Strybel, T. Z. (2015). Using eye movements to test assumptions of the situation present assessment method. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9173*, 45–52.

Sulaiman, S., Rohaya Awang Rambli, D., Fatimah Wan Ahmad, W., Hasbullah, H., Oi Mean, F., Nordin Zakaria, M., Kim Nee, G., & Rokhmah Shukri, S. M. (n.d.). Volume (3) : Issue (1). In *International Journal of Computer Science and Security (IJCSS)* (Issue 3).

Suleri, S., Shishkovets, S., Pandian, V. P. S., & Jarke, M. (2019). Eve: A sketch-based software prototyping workbench. *Conference on Human Factors in Computing Systems - Proceedings*, 1–6.

Suziah, S., Rambli, D. R. A., Ahmad, W., Fatimah, W., Hasbullah, H., Foong, O. M., Nordin, Z. M., Goh, K. N., & Sukri, S. R. M. (2009). Asking Users: A Continuous Usability Evaluation on a System Used in the Main Control Room of an Oil Refinery Plant.

International Journal of Computer Science and Security, *3*(1), 34–42. http://eprints.utp.edu.my/1167/

Tatler, B. W., & Vincent, B. T. (n.d.). *IN PRESS MANUSCRIPT VERSION The prominence of behavioural biases in eye guidance*.

Teikari, P. (2007). Automated pupillometry. *Helsinki University of Technology, Centre for Metrology …*, 1–35. http://users.tkk.fi/~jteikari/Teikari_AutomatedPupillometry.pdf

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

Tole, J. r., A.T., S., Vivadou, M., Ephrath, A., & Young, L. R. (1983). *Visual Scanning Behaviour and Pilot Workload* (Vol. 3).

Tran, T. Q., Boring, R. L., Dudenhoeffer, D. D., Hallbert, B. P., Keller, M. D., & Anderson, T. M. (2007). Advantages and disadvantages of physiological assessment for next generation control room design. *IEEE Conference on Human Factors and Power Plants*, 259–263.

Tran, T. Q., Boring, R. L., Dudenhoeffer, D. D., Hallbert, B. P., Keller, M. D., & Anderson, T. M. (2007). Advantages and disadvantages of physiological assessment for next generation control room design. *IEEE Conference on Human Factors and Power Plants*, 259–263. https://doi.org/10.1109/HFPP.2007.4413216

Uebelbacher, A. (2014). *The fidelity of prototype and testing environment in usability tests*.

Ulrich, T; Boring, R; Lew, R. (2018). QUALITATIVE OR QUANTITATIVE DATA FOR NUCLEAR CONTROL ROOM USABILITY STUDIES? A PRAGMATIC APPROACH TO DATA COLLECTION AND PRESENTATION. *Proceedings of Human Factors and Ergonomics*, 1674–1678.

van de Merwe, K., van Dijk, H., & Zon, R. (2012). Eye Movements as an Indicator of Situation Awareness in a Flight Simulator Experiment. *International Journal of Aviation Psychology*, *22*(1), 78–95.

Van Der Meer, E., Beyer, R., Horn, J., Foth, M., Bornemann, B., Ries, J., Kramer, J., Warmuth, E., Heekeren, H. R., & Wartenburger, I. (2010). Resource allocation and fluid intelligence: Insights from pupillometry. *Psychophysiology*, *47*(1), 158–169.

van Lankveld, G., Sehic, E., Lo, J. C., & Meijer, S. A. (2017). Assessing Gaming Simulation Validity for Training Traffic Controllers. *Simulation and Gaming*, *48*(2), 219–235.

Van Leijen, M. (2018). *Traffic Volumes on Dutch freight lines.* Railfreight.com Online Magazine. 15.

Vidulich, M. A. (2000). The relationship between mental workload and situation awareness. *Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Association, "Ergonomics for the New Millennium,"* 460–463.

Vidulich, M.A. & Tsang, P.S. (2012). Human Factors Engineering and Systems Design. In G. Salvendy, *Human Factors Handbook* (pp.243-273). Hoboken, New Jersey: John Wiley and Sons.Walker, M., Takayama, L., & Landay, J. A. (2002). High-Fidelity or Low-Fidelity, Paper or Computer? Choosing Attributes when Testing Web Prototypes. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *46*(5), 661–665.

Wanyan, X., Zhuang, D., & Zhang, H. (2014). Improving pilot mental workload evaluation with combined measures. *Bio-Medical Materials and Engineering*, *24*(6), 2283–2290.

West, R., & Lehman, K. R. (2006). Automated summative usability studies: An empirical evaluation. *Conference on Human Factors in Computing Systems - Proceedings*, *1*, 631–639.

WHITEFIELD, A., WILSON, F., & DOWELL, J. (1991). A framework for human factors evaluation. *Behaviour & Information Technology*, *10*(1), 65–79.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*(3), 449–455.

Wikström, J., Kauppi, A., Andersson, A. W., & Sandblad, B. (n.d.). *Designing a graphical user interface for train traffic control*. 26.

Williams, B., Quested, A., & Cooper, S. (2013). Can eye-tracking technology improve situational awareness in paramedic clinical education? *Open Access Emergency Medicine*, *5*, 23–28.

Wilson, J. R., Ryan, B., Schock, A., Ferreira, P., Smith, S., & Pitsopoulos, J. (2009). Understanding safety and production risks in rail engineering planning and protection. *Ergonomics*, *52*(7), 774–790.

Wilson, J. R., Ryan, B., Schock, A., Ferreira, P., Smith, S., & Pitsopoulos, J. (2009). Understanding safety and production risks in rail engineering planning and protection. *Ergonomics*, *52*(7), 774–790.

Wood, K. L., Jensen, D., Bezdek, J., & Otto, K. N. (2001). Reverse engineering and redesign: Courses to incrementally and systematically teach design. *Journal of Engineering Education*, *90*(3), 363–374.

Yin, X., Yurcik, W., Li, Y., Lakkaraju, K., & Abad, C. (2004). VisFlowConnect: Providing security situational awareness by visualizing network traffic flows. *IEEE International Performance, Computing and Communications Conference, Proceedings*, *23*, 601–607.

Young, M. S., & Stanton, N. A. (2002). Attention and automation: New perspectives on mental underload and performance. *Theoretical Issues in Ergonomics Science*, *3*(2), 178–194.

Young, M. S., Stanton, N. A., & Walker, G. H. (2006). In loco intellegentia: human factors for the future European train driver. *International Journal of Industrial and Systems Engineering*, *1*(4), 485–501.

Zargari Marandi, R., Madeleine, P., Omland, Ø., Vuillerme, N., & Samani, A. (2018). Eye movement characteristics reflected fatigue development in both young and elderly individuals. *Scientific Reports*, *8*(1).

Zhang, C. (2016). The Effect of Situation Presence Assessment Method (SPAM) on Air Traffic Control Students' Workload and Performance in High-Fidelity Simulations. In *Journal of Chemical Information and Modeling* (Vol. 110, Issue 9). Arizona State University.

Zimmerman, J., Forlizzi, J., & Evenson, S. (2007). *Research Through Design as a Method for Interaction Design Research in HCI*. 493–502.

Zimmerman, J., Forlizzi, J., & Evenson, S. (2007). Research through design as a method for interaction design research in HCI. *Conference on Human Factors in Computing Systems - Proceedings*, 493–502.

Zimmerman, J., Forlizzi, J., & Evenson, S. (2007). Research through design as a method for interaction design research in HCI. *Conference on Human Factors in Computing Systems - Proceedings*, 493–502.

Appendix

Table 1. Eye tracking Measures and SA and MWL in Safety Critical Research Environments

(workload, SA,)
(*= significant result, levels of SA)

Reference	Scanpath	Blink rate	Blink Frequency	Fixation rate / proportion (in given time period)	Fixation frequency (in AOI)/ refixations	Pupillometry	Gaze/ scan Entropy	Saccade Amplitude	Dwell (Gaze duration in AOI)	Other Measures
Allsop et al. (2017)							**X**			
Argyle et al. (2020)				**X**	**X**				**X**	Freeze probe (SA)
Bjorneseth et al.(2012)	**X**	**X**			**X**			**X**		
Bracken et al. (2019)					**X**					SPAM, Bedford WL scale, fNIRS sensor
Buettner et al. (2018)						**X***				Random forest method, Subjective perceived task difficulty
Coyne & Sibley (2015)				**X***			**X (proportion of movement)**			Position based freeze probe,
De Greef et al.(2009)						**X***		**X**	**X***	
De Winter (2018)				**X**					**X**	
Desvergez et al. (2019)					**X**		**X (heatmap)**		**X**	SAGAT
Di Flumeri et al. (2018)					**X**					EEG, NASA TLX
Diaz-Piedra (2019)*							**X**			
Di Stasi et al. (2009)					**X**	**X**		**X**	**X**	Peak saccadic velocity*, saccadic duration, Ergonomics evaluation q, Mental workload test*
Di Stasi et al. (2010)								**X**		MWT, saccadic peak velocity* and duration
Fabio et al. (2015)							**X**			Time before correct fixation

Foy & Chapman(2018)							**X (horizontal spread of search)***		**X**	Nasa TLX, HRV, respiration skin conductance*
Gartenburg, Curry & Trafton (2011)				**X obj fixations**	**X**				**X Faster dwell**	Compared re- and novel-fixations
Ha et al. (2006)		**X**	**X**		**X**					Nasa TLX, Cooper Harper scale
Hareide, Ostnes & Mjelde (2016)	**X**			**X**	**X (fixation count)**				**X (saccade & fixations)** **// X (fixation)**	Entry time (duration first fixation), hit ratio (how many subjects looked into AOI), revisitors (how many subj. Looked into AOI>2)
Hareide & Ostnes (2017)	**X**			**X**	**X**			**X**	**X**	Backtracks
Hazansadeh et al., (2018)	**X**			**X**				**X**	**X**	Fixation time (how quickly fixates on an AOI), SART
Hooge & Camps (2013)^							**X**	**X**	**X**	Arrow plot (avg scan path), T50 (time until half of P's reach first target AOI)
Ikuma et al. (2014)				**X ***						NASA TLX, SWAT, SAGAT, mouse-tracking
Iqbal et al. (2004)						**X**				"Subjective MW ratings", average percentage change of pupil size
Katrahmani, Ahmadi & Romoser (2017)							**X (L3-strategic glancing patterns)**		**X (L2)**	SA verbal protocol (3 level). Glancing (L1)- used VP w ET to assess levels
Kearney et al. (2018)					**X**				**X**	Saccadic velocity, COOPANs acoustic alert, semantic alerts
Kovedsi et al. (2018)	**X**	**X**	**X**			**X**			**X**	Matrix density, scanpath length and duration, blink duration, verbal protocol
Li, Chiu & Wu (2012)									**X***	NASA TLX, (performance was SA)

Lu, Coster & De Winter (2017)										Self report self sufficiency and task difficulty
Mark et al. (2020)						**X***		**X***	**X***	Cognitive testing, fNIRS, EEG, ECG, EOG, PPG, SA= task difficulty
McIntre et al (2014)						**X**				Pupil velocity
Miramonte (2017)				**X***	**X***					SPAM
Molina, Retamal & Lira (2018)						**X**				(electrodermal activity (EDA), electrocardiogram, photoplethysmo-graphy (PPG), electroencephalogram (EEG), temperature
Moore & Gugerty (2010)				**X***					**X**	Spam/ SAGAT hybrid, “scene fixations” (entropy?), NNI
Noah & Rothrock (2015)		**X**	**X**			**X***				NASA TLX*, SAGAT like probe
Nocera et al. (2016)							**X**			AOI transitions, Nasa TLX, Cognitive Failures Q.
Othman & Romli (2016)						**X***				Nasa TLX,
Orlandi & Brooks (2018)-						**X**				ECG< Nasa TLX, MWL likert scale subj rating
Petersen et al. (2019)				**X***						SART, HRV, trust in automation, (monitoring frequency switching AOIS)
Pfleging et al. (2016)						**X***				NASA TLX, varying light conditions
Recarte et al. (2008)*		**X***				**X**				NASA TLX
Schulz et al. (2011)-						**X***		**X**	**X***	HRV*, Rating of Perceived Exertion
Schwalm, Keinath & Zmmer (2008)						**X***				Nasa TLX, Index of Cognitive Activity
Schwerd & Schulte (2020)					**X**					SPAM
Stapel, Hassnaoui & Happee (2020)				**X**	**X**			**X**	**X (glance)**	SAGAT like recognition based method, first saccade

										angle, duration preceding fixation
Sturre et al.(2015)				**X**	**X**				**X**	SPAM
Szewczyk et al. (2020)									**X**	Gazeplots, heatmaps, saccade frequency*
Tokuda et al. (2011)								**X***		"Subjective MWL rating"
Van De Merwe (2010)				**X**			**X**		**X**	CARS, subjective SA rating
Wanyan, Zhuang & Zhuang (2014)		**X***				**X***				NasaTlx, ERP*, ECG
Zhang et al. (2017)						**X* (correlations significant for fatigue)**				Surface electromyography (sEMG) signals of the bilateral trapezius, bicipital, brachioradialis and flexor carpi ulnaris (FCU) muscles
Zheng et al. (2012)		**X***	**X***							Nasa TLX

Q.= Questionnairre
Subj.=Subjective
w= with
^=didnt look at SA or MW, but interesting enough to be included

Table 2. Situation Present Assessment Method Usage in Research Environments

Reference and Field	Test Scenarios	SPAM Qs (Number, modality, frequency)	SPAM results (Average Reaction Time, latency)	Other Methods and results
Bracken et al. (2019)	1.5 hour testing sessions, as well as information sessions about SA and WL and task (status check task, with secondary tasks such as vehicle launching)	2= n/a M=yes or no, or multiple choice F= n/a	n/a	Eye tracking (AOIs), Bedford WL scale, self report SA (likert four point, HbO (prefrontal oxygenated blood volume) and cardiac information during activities with low physical activity, System Usability scale) -One condt. More fixations (suggesting lower WL), all condt.s has different usability scores
Durso (1999)	SPAM in one of three scenarios (one other with SAGAT and one baseline)	N= 6 M= Occurred when landline was called (relevant to task) F= n/a	L=~4 seconds -Better outcome when fewer remaining activities at end, predicted subject matter expertise well	-Nasa TLX, SART and SAGAT -SAGAT=lower % correct than SPAM -TLX high intercorrelations among items
Durso et al. (2006)	-SPAM in two of six scenarios (two more had offline queries and two were control) -Six memory span tasks for air traffic control, whereby accuracy indicated performance	N= 6 per scen. M=half heard with earphones, half read in lower right corner -> responses binary, 2 or 8 options, all had warning F=first 3-4 min of scen, after every two min	RT=0-30 seconds (based on that queries were presented every 2.53 mins depending on operator activity) L=30 s- 1 minute -Sig. Better at predicting performance than offline, high incremental validity, poorest for level 3 SA	-Cognitive, personality and demographic variables, and an offline query -Variables had lower incremental validity than SPAM, offline query had a high correlation (logical, meant to measure same construct)
Miramontes (2017)	-SPAM in all scenarios, had queries differing in specificity (general aircraft questions vs specific ones) and priority (high was regarding conflicts or position, whereas low was regarding an aircraft not in action)	N= 12/ scenario M= binary response, differed in specificity for subj. matter, ready button for operator and auditory alert F=every three minutes	RT=indicated WL L=indicated SA Both were significantly longer for higher specificity queries	-Four workload queries also -Eye tracking: fixation frequency and duration at target AOI,were more likely to fixate on the display during higher WL, scenario and priority did not have a sig. Effect on the
Pierce et al. (2008)	20 minute scenarios, with first query occurring at three minutes, used a ready probe (auditorily), air traffic scenarios test context	N= 7 M=verbal response, indicated ready by pressing space bar F=every 2.83 minutes	RT=measured from ready signal onset, used also for list and shadowing conditions -similar outcomes to TLX (highest in list condt.) -In three of the outcomes SPAM conditions had sig poorer performance than baseline q.	-word shadowing and list recital (as separate conditions to SPAM), and two questionnaire conditions (low fidelity) -Nasa TLX (both SPAM and shadowing ranked lower than list recital for task workload)
Pierce (2012)	First probe presented four mins into scenario	N= 6 M=Probe presented auditorily, indicated ready by pressing spacebar F=2.83 mins	RT=Measured from onset, length of time not stated	Nasa TLX, word shadowing, list and two baseline -Same outcomes at Pierce 2008
Schwerd & Schulte (2020)	2 scenarios with training	M= threat node would be	RT=3 seconds (to report	-Eye tracking (used

	provided, SPAM like probes occurred when threats were present and also for changes in conditions	activated, had to be reported M= also measured time until pilot reported altitude as an adapted SPAM (for changing conditions)	the presence of the threat node for it to be correct) -3 times it took over six seconds -Found that neither of the SPAM methods worked well since operators were not fond of frequent updates	fixations to indicate contextual interface features- found that it supported SA investigation) and SART (as a baseline SA measurement)
Strybel et al. (2008)	20 minutes scenario, four scenarios whereby pilots could self pace, with an easy and a hard condition, and two with SAGAT and two with SPAM	N=10 queries M=computer based visual query (datalink with MACS software) F=every two minutes	L=predicted the indicated airspeed (IAS) and thus interfered with the scenario, however accuracy was still better in SPAM than SAGAT scenarios	Nasa TlX and SART at the end of every scenario, as well as SAGAT as a comparison (which was found to be more intrusive in subjective pilot ratings)
Sturre et al. (2015)	Four forty min scenarios ten min breaks between, low and high priority probes and general and specific probes	N= 12 M=Presented on a touch screen, given a ready prompt F=3 minutes	RT=N/a L=N/A	Eye tracking (looked at fixations within target AOIs)
Zhang (2016)	Two scenarios, one with SPAM and one without	N= 8 total (four per scen) M=asked verball F=every 3 mins	L=4 seconds No significant differences in errors found between conditions	Nasa TLX (no found to be significantly different between conditions)

Table 3. Eye-tracking Metric Definitions

Metric Name and Definition	References	Notes
Movement direction measures (direction of movements in stim space, also rel. to vision)		
Anti-saccadic metric Focusses on whether operators can inhibit / reverse saccadic direction relative to a suddenly appearing/ distractor target.	(Holmquist & Nystrom, 2011)	This metric is made-up of six sub-measures
Saccadic direction To determine whether one of three of the following saccadic movements occurs: Backtrack/lookback (saccadic movement in the opposite direction of previous saccade- counting events), lookahead (saccadic direction moves towards upcoming target AOI) or leading saccade (which is related to upcoming eye-movement and shown to occur when the saccade skips over the target and slows down in order for the target to catch up with the gaze)	(Holmquist & Nystrom, 2011)	Lookbacks are dependent upon the operational definitions Lookaheads occur in anticipation of event occurrence, and fixation occurs before AOI, and are shown to be less common than guiding fixations, and take less time
Scanpath direction through AOI Calculable using the first and final fixations within an AOI coordinates during a defined period, in order to determine the gaze orientation of operators	(Renshaw et al., 2003*)	Need to relate it back to task context and search strategies to give it proper meaning
Movement amplitude measures (in terms of displacement or velocity)		
Saccadic amplitude Used to determine how far the gaze was displaced (in pixels),with shorter ones indicating focussed inspection of details (as opposed to scanning) or a high frequency of visual information presentation, as well as increased cognitive load and difficulty of task (in terms of searching)	(Philips & Edelman, 2008; Tatler & Vincent, 2008; Recarte & Nunes, 2003)*	Can differ per eye, need to be mindful for that.
Scanpath length the sum of all the saccadic amplitudes in a scanpath (usually calculated in Euclidean distance by algorithms)	(Holmquist & Nystrom, 2011)	
Saccadic duration the time between two smooth pursuits or fixations, that is thought to represent a period of no visual input	(Holmquist & Nystrom, 2011)	Associated with increased levels of fatigue or task difficulty
Scanpath duration the sum of all fixation durations throughout a task or task subsection	(Holmquist & Nystrom, 2011)	Myriad of operational definition, the one here is chosen due to simplicity
Blink duration the entire time taken to complete a blink from when the eye begins to close to when it opens again, and is thought to be proportional to task workload and inversely proportional to state of vigilance.	(Kovedsi et al., 2017 Noah & Rothrock, 2015, Marquart et al., 2015)	
Saccadic velocity calculated as the average saccadic velocities in a specific period of time, with lower saccadic velocity being associated with decreased vigilance, or target	(Holmquist & Nystrom, 2011)	

predictability (when an anticipatory saccade move in the direction prior to target onset- and have been found to be slower than reactionary saccades)		
Pupil velocity The speed at which the pupil expands and contracts in a specified period of time (with expansion and contraction being measured separately since constriction is faster)	(Holmquist & Nystrom, 2011; Buettner et al., 2018)	Has been associated with increased vigilance and higher workload
AOI and transition movement metrics		
Order of AOI entries Assessed by creating a table ordering the participants AOI entries (ranked chronologically), and average each AOIs value in order to ascertain the average order of AOI entrance to show which AOIs are perceived first.	(Holmquist & Nystrom, 2011)	Variation between rankings represents the tendency to visit similar AOIs between participants) and can be used to calculate kendall's coefficient of concordance.
Transition matrices:	Hooge & Camps, 2013; Blascheck et al., 2016	Two common methods of measuring are either Create transition cells and make an arrow plot (can also represent entropy) OR use transition trees (see how AOIs nested)
AOI transitions The frequency of transitions between AOI's within a set period of time. Shows how shifts in gaze represents which AOIs are most salient at which point in time, and how AOIs are related to each other.	Kovedsi et al.(2017), Holmquist & Nystrom (2011)	
Scanpath Entropy The degree of gaze guidance, with a high degree occurring when the majority of participants (or operators, in the case of the control room) have similar scanpaths.	Hooge & Camps, 2013	Allows investigation into whether the system/ interface supports visual search efficiency by assigning a target AOI in a top-down manner
Gaze Entropy Uses an entropy calculation to represent the randomness of gazing patterns.	(Van De Merwe et al., 2012; Moore & Gugerty, 2010, Diaz-Piedra, 2019; Allsop et al. (2017);Coyne & Sibley (2015);Desvergez et al., 2019; Fabio et al., 2015; Othman & Romli, 2016)	Higher entropy rates in the case of attending to an abnormality resulted in fewer and more random fixations as well as higher error margins, and that teams with more structured gaze patterns performed much better. Higher SA being associated with more systematic viewing patterns within a team
T50 measure The time at which half of the operators would find the target/ salient AOI or AOI subsection at a crucial moment.	Hooge & Camps, 2013	Allows for comparison between interfaces and interface features during crucial task moments (with smaller T50 equating to better VSE support)
Position measures (stillness of gaze in one or more positions)		
Fixation the average x,y coordinates from data falling within a specified space over for a minimum of a specified time, when the gaze is fixed.	(Holmquist & Nystrom, 2011)	Looking at basic fixations within an AOI makes it possible to determine for example where location, direction, guidance and checking occurs
Dwells Dwells have the whole area of the AOI as their position value, since they are an event, and are categorical variables in nature	(Holmquist & Nystrom, 2011;Bergstrand, 2008; Van De Merwe et al., 2010 Hasanzade et al., 2016)	Must be careful AOI isn't too large, to avoid inaccuracy

Dwell duration the time period over which the dwell occurs (in ms), that is how long is attention fixated within an entire AOI	(Holmquist & Nystrom, 2011)	Has been shown to represent uncertainty and poor SA in novice operators (with longer duration), but also has been shown to represent interest (rooted in task context)
Fixation duration Differs from dwell duration in that it doesn't represent the lingering of gaze within an entire AOI for a block of time, but rather shows a more precise pinpoint of where attention was focussed over time n.b. Is thought to represent a variety of cognitive processes in different contexts	(Argyle et al., 2020; De Winter, 2018; Holmquist & Nystrom, 2011)	Longer durations represent: -deeper level processing and environmental adaptation (in novel situations) -lowered vigilance in very familiar settings, in which stimuli is low, or difficulty in extracting information -complexity of information presented -prolonged (on only salient areas) when expertise is higher Shorter durations are shown to indicate higher operator stress.
Fixation density How many fixations are occurring within a given field Representation of the spread of fixation clusters in a specific period of time- represents searching when not dense? Counts the number of fixations and divides it up over the area in which it has occurred- High MWL	(Holmquist & Nystrom, 2011)	Most simplistically represented either visually or with NNI - nearest neighbour index (in pixels)
Glissadic aftermath duration The slow, drifting eye movement at the tail end of the saccade as it turns into the fixation, which could also represent visual intake, and be important in determining the true fixation duration period.	(Holmquist & Nystrom, 2011)	Usually missed by event algorithms
Skewness of fixation frequency distribution Examines whether the fixations tend towards long or short in a fixed trial or time period	(Holmquist & Nystrom, 2011)	
Pupil diameter The diameter of the pupil at particular snapshots throughout the task time, represented through pixel values.	Singh et al., 2017; Piquado, 2010; McIntre et al., 2014;Iqbal et al., 2004; De Greef et al., 2009; Mark et al., 2020, Molina, Retamal * Lira, 2018)	Represents mental workload and anticipation (larger) and fatigue (smaller), however it can be very sensitive to fluctuating lighting conditions
Numerosity measures		
Frequency measures Can be applied to all key metrics () and come in the format of spatial (i.e. how many in one area) versus temporal (i.e. how many in one time period)- and are again grounded in the task context	(Holmquist & Nystrom, 2011)	Temporal frequency measures are not to be confused with rate measures- rate is how many occurred per fixed unit of time, whereas temporal frequency is how many occurred during a specific period.
Glissadic proportion The number of glissadic eye movements occurring per saccade.	(Holmquist & Nystrom, 2011)	Shown to be associated with vigilance
Saccade numerosity measures Frequency (from trial start to end), rate (the number per second, during a set period) and proportion (the number in a subgroup, divided by the entire amount)	Kovedsi et al.(2017), Holmquist & Nystrom (2011)	
Blink rate and frequency	(Noah & Rothrock, 2015, Marquart et al., 2015; Holmquist & Nystrom,	Both are associated with mental workload. Must be careful that data taken from people wearing

number of blinks occurring per minute (usually), and has been shown to increase with mental workload and fatigue- and are not a continuous measure. Blink frequency also because it has been found to increase steeply at the beginning of an increase in workload, but remain steady throughout the later part of a task (which is thought to indicate that some kind of orientation phase could occur).	2011;Bjorneseth et al., 2012; Ha et al., 2006; Kovedsi et al., 2018; Noah & Rothrock, 201)	glasses doesn't cause Smarteye to interpret glinting (and thus a loss of pupil data) as a blink, also need to be aware of individual differences which are common with this measure.
Inter-blink interval The amount of time occurring between blinks in seconds.	(Holmquist & Nystrom, 2011)	Calculated as the inverse of the blink rate.
Fixation numerosity measures frequency (from trial start to end), rate (the number per second, during a set period) and proportion (the number in a subgroup (for the current context, one AOI), divided by the entire amount)	(Argyle et al., 2020, Holmquist & Nystrom, 2011; Coyne & Sibley, 2015; Bracken et al., 2019, Di Flumeri et al., 2018)	Frequency has been shown to be be negatively related to search efficiency and level of experience, and positively related to interest and rate to be negatively correlated with scenario difficulty, and correlated to performance)
Dwell numerosity measures Frequency- usually measured as the amount of dwells in each AOI,two consecutive dwells in the same AOI being allowed (unless there is whitespace in the stimulus), also common to look at proportion in a specific time period and rate.	(Holmquist & Nystrom, 2011)	Signifies increased practice (a lower amount), and semantic importance and complexity of features (associated with more).
Proportion measures Prop of participants that look at an AOI- to flesh out which AOIs are missed and to detail whether some AOIs that might have long dwell times are being missed by some participants- has been shown to signify contextual predictability Proportion of trials: looks at the proportion of trials in which certain behaviours occurred, such as proportionally how many fixations occurred in a particular trial	(Holmquist & Nystrom, 2011)	Inspect the amount of times a specific criterion is satisfied in order to give more dimensionality to eye-tracking analysis in the context of the experimental design.
Number of re-fixations In total always equals the number of dwells minus one, but can instead be done as a proportion also (whereby the number of re-fixations are divided by all the targets which have been fixated upon) to show which targets required more attention	(Holmquist & Nystrom, 2011)	Could also make a threshold for refixation time limit, that is, all re-fixations which occurred within thirty seconds
Fixation to importance ratio: measurable by looking at individuals as opposed to cohorts average gaze duration and number of fixations in specific AOIs	(Kovesdi et al., 2015; Ha & Seong, 2007)	The manner in which an operator allocates their fixation to an AOI in terms of its degree of importance to the task (as determined by group averages)
Latency (time delay) and distance (from one point to another) measures:		
Eye-mouse distance Expressed in pixels, and conceptualised as the distance between the point of gaze and the mouse.	(Holmquist & Nystrom, 2011)	Used to signify to what extent mouse-tracking can be used as a substitute for eye-tracking.
Saccadic gain Measured as the distance between the ending point of the saccade and the beginning of the target, and is expressed as a percentage.	(Holmquist & Nystrom, 2011)	100% equates to landing in the target (and less being under (undershoot) and more being over (overshoot) when the measure is expressed simplistically.

Saccadic latency The time after the onset of the target for which it takes for the saccadic movement to being, thought to signify reaction time	(Holmquist & Nystrom, 2011)	Shown to be positively associated with split attention (distractors/ targets in multiple parts of the visual field) and negatively associated with anticipation and predictability of appearing targets, and measured in ms.
Smooth pursuit latency The measure of reaction time for targets which move in a smooth manner, have been also shown to be negative in the case of anticipation	(Holmquist & Nystrom, 2011)	Difficult to calculate onset, and might need to be done manually and by looking at individual trials.

Questionnaire 1. Post testing Questionnaire.

Overall background information of Participants (not to be used individually)

1. How much work experience do you have as a train traffic controller?

……

2. How much work experience do you have in the railway sector?

……

	Completely disagree				**Completely Agree**	
1. The second scenario was easier because of prior experience with the first scenario	1	2	3	4	5	w.n.

Usability

Please indicate to what extent you agree with the beneath statements.

1 = Helemaal mee oneens; 2 = Oneens; 3 = Niet eens/niet oneens; 4 = Mee eens; 5 = sterk eens; w.n. = niet van toepassing/weet niet

	Helemaal mee oneens				**Helemaal mee eens**	
1. I could easily acclimatise to the simulator.	1	2	3	4	5	w.n.
2. I could easily acclimitise to the new service area.	1	2	3	4	5	w.n.
3. I could easily acclimitise to the ERTMS UI.	1	2	3	4	5	w.n.
4. I would gladly use the ERTMS interface regularly.	1	2	3	4	5	w.n.
5. I think the ERTMS interface is unnecessarily complex.	1	2	3	4	5	w.n.
6. I think the ERTMS interface is easy to use.	1	2	3	4	5	w.n.
7. I require support from a technical person to use the ERTMS interface.	1	2	3	4	5	w.n.
8. I think the various functionalities in the ERTMS interface are very well integrated.	1	2	3	4	5	w.n.
9. I feel that there are too many contradictions in the ERTMS interface.	1	2	3	4	5	w.n.
10. I could imagine that most ERTMS users can learn quickly to use the interface.	1	2	3	4	5	w.n.
11. I find the ERTMS interface cumbersome to use	1	2	3	4	5	w.n.
10. I feel very familiar with the ERTMS interface	1	2	3	4	5	w.n.
11. I had to learn a lot before I could use the ERTMS interface easily	1	2	3	4	5	w.n.

Simulator Validity

	Completely agree				**Completely disagree**	
1. I found the timetable display in the current simulation task to be completely realistic.	1	2	3	4	5	w.n.

2.The simulation environment felt pretty much like my own workplace	1	2	3	4	5	w.n.
3.I found the infra model in the current simulation task to be completely realistic	1	2	3	4	5	w.n.
4.I couldn't find all the necessary information in the simulator in order to complete my task well	1	2	3	4	5	w.n.
5. The train service process corresponds process wise to real train service processes	1	2	3	4	5	w.n.
6. The simulator consists of necessary functions that are required to complete this task	1	2	3	4	5	w.n.
7.The chosen scenario in the simulator represents a situation that is encountered in the railway sector	1	2	3	4	5	w.n.
8. The information that can be accessed through information sources in the simulator can be altered in the same manner as in an actual interface	1	2	3	4	5	w.n.
9. The processes (interactions, communications) in the simulator match with processes in a comparable situation in my workplace	1	2	3	4	5	w.n.

Intent and goal of game session

Completely disagree					**Completely agree**	
1. The objective of the session was obvious	1	2	3	4	5	w.n.
2. I can see the added value of this research	1	2	3	4	5	w.n.
2. The instruction material was sufficient for preparation for the scenarios in the experiment.	1	2	3	4	5	w.n.

Do you still have observations?

Once again, many thanks for your participation!