Human Performance in Rail:
determining the potential of physiological data from wearable technologies

Abigail C Fowler, MSc. BSc. (hons.)

Human Factors Research Group, Faculty of Engineering and
Horizon Centre for Doctoral Training, Computer Science

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

March 2023

Abstract

This research focuses on how personal data from wearable physiological measures can be used to assess the Mental Workload (MWL) of staff in the rail industry. Automation technologies are being implemented in the rail industry to improve operational performance and capacity. These new technologies are changing the role of staff. This research considers how temporal physiological data present an opportunity to supplement existing workload assessment methods to measure the impact of these technology changes. The research explores how wearable physiological measures could be applied in live operations to collect real-time data with minimal task interference. Whilst the research focuses on railway signallers, the research has implications for other roles in the rail industry and other industries.

The research included three studies and two literature reviews. The initial industry interview study identified the benefit of more continuous data to assess human performance, including successful performance. A detailed review of candidate technologies was then performed solely on physiological measures to extend the knowledge in this area. To assess the potential of physiological measures to provide this continuous data, a simulation study of railway signalling tasks was conducted with an Electrodermal Activity (EDA) wrist strap for alertness and stress and a Heart Rate Variability (HRV) chest strap for uncertainty and increased MWL. The limited application of these measures in rail research provided a suitable research gap for the research to pursue.

The simulation study found physiological data provided visibility of individuals' personal experience of workload. The interplay of EDA, HRV, task demand and subjective workload over time were visible in the storyboard for each participant. The simulation study provided two key contributions to the thesis. Firstly, EDA identified moments in workload during the task, indicating moments of realisation, and periods of uncertainty, or strain due to time

pressure. Such data could be used in staff debriefs to better understand their workload, and tailor training. Secondly, average HRV had a strong negative correlation with average subjective workload. HRV could provide a real time indicator of workload and provide visibility of staff effort to managers.

The final study was an interview and survey study of staff perspectives on the potential use of these measures. This study replaced a live trial which could not proceed during COVID-19 related restrictions. The study found wearable devices suit use in the live operational environment, with the wrist strap rated the most suitable due to low distraction. Trust emerged as a key factor for staff to accept the use of wearables, particularly if named data is shared. Tangible benefits that lead to improvement in operations was identified as one way to build this trust.

An additional contribution of the thesis, drawing on all studies and literature reviews, was to propose a new theoretical perspective on MWL, based on physiological data. A Novelty of Events and Autonomic State (NEAS) model is proposed as a preliminary conceptualisation. It shows how individuals may vary in the impact workload has on their performance and how physiological data may be used to identify this. The concept of Novelty of Events includes aspects of tasks that an individual has not performed before, including those introduced by new technology or procedures. The NEAS model suggests how support in the form of tailored training, or shift breaks, could be used to support improved human performance. Following on from this thesis, a priority for further empirical work would be to trial EDA using a wrist strap that uses a repeated measures approach to determine to what extent individual physiological data changes over time.

Acknowledgements

My thanks go firstly to my family Graeme, Alex, Charlotte, my parents, in-laws, and Martha. They have encouraged and supported me throughout, and without their support I could not have embarked on a PhD and would not have completed one. Secondly to my supervisors Cath Harvey, Max L Wilson, Sarah Sharples, and Dave Golightly for their guidance and belief in me throughout. My thanks also go to my industry sponsors including Huw Gibson at RSSB, Chris Hack and Mike Carey at Network Rail, and staff at Network Rail and various Train Operating Companies and Freight Operating Companies who enabled my industry studies to go ahead.

The companionship of fellow PhD candidates and researchers has provided invaluable support throughout. This included colleagues in the Horizon Centre for Doctoral Training (CDT), the Human Factors Research Group, and the Brain Research Group, all at the University of Nottingham. A particular mention goes to Adrian Marinescu and Horia Maior who piqued my interested in physiological measures early on and taught me how to use the equipment along with Paul Tennent. I would also like to thank University professional support staff, in particular Lee, Donna in Engineering and Andrea, and previously Emma, in Horizon CDT.

Friends have been my moral boosters Karen, Chloe, Emily, and Ayesha who generously listened to me talk about my research, Jen for the quiet place to write when it was most needed, and all my proofreaders. In addition, Kermode and Mayo who helped keep me going with their reassuring Wittertainment, to the NHS staff who saved my life, and Kinnerton's for my favourite gluten free dairy free chocolate.

Finally, author Douglas Adams for highlighting the importance of understanding both the answer and what the question means.

Publications

One publication has been completed during the period of this research: Fowler, A., Golightly D., Sharples, S., Harvey, C., Wilson M. L., and Gibson, H. (2019) Human performance in rail: current assessment and future opportunities in Charles, R. and Golightly, D. (Eds) *Contemporary Ergonomics and Human Factors 2019*. ISBN 978-1-9996527-1-5

Table of Contents

List of Figures

List of Tables

Glossary

ANS	Autonomic Nervous System
ARS	Automatic Route Setting
ATC	Air Traffic Control
ATO	Automatic Train Operation
ATWS	Automatic Train Warning System
ASWAT	Adaptive Subjective Workload Technique
AWS	Automatic Warning System
Bogie	A wheeled undercarriage pivoted below a rail vehicle
CCF	Control Centre of the Future, shows the real time movement of trains
CCTV	Closed Circuit Television
DRS	Dynamic Route Setting
DSD	Driver's Safety Device
DVD	Driver's Vigilance Device
EDA	Electrodermal Activity
EDR	Electrodermal Response
EMCC	East Midlands Control Centre
ERTMS	European Rail Traffic Management System
FOC	Freight Operating Company
GDPR	General Data Protection Regulation
GSM-R	Global System for Mobile Telecommunications for Railways
GSR	Galvanic Skin Response
Head Code	Term for the Train Reporting Number. A unique code to identify a train, including class of train
HRV	Heart Rate Variability
IECC	Integrated Electronic Control Centre
IWS	Integrated Workload Scale
Level Crossing	A place where a railway and a road cross at the same level
Lever Frame	In signal boxes, pulling the lever mechanically operates points. Used since 19th century.
LOM	Local Operations Manager
MWL	Mental Workload
NASA TLX	National Aeronautics and Space Administration Task Load Index
NEAS model	Novelty of Events Autonomic State model
NX panel	Entry Exit Panel, a type of signalling control system
ORR	Office of Rail Regulation
OTDR	On Train Data Recorder
Panel	Signaller's 'workface' to control area and set routes
PNS	Parasympathetic Nervous System
Points	A junction of two railway lines that can be set to guide a train onto alternative routes
Possession	When engineers are responsible for sections of track during maintenance

RAIB	Rail Accident Investigation Branch
RBTNA	Risk Based Training Needs Analysis
RSSB	Rail Safety and Standards Board
Rule Book	Book of rules for the safe operation of the railway network
SARS	Signaller's Assistant Route Setting
SCL	Skin Conductance Level
SCR	Skin Conductance Response
SNS	Sympathetic Nervous System
SPAD	Signal Passed At Danger
SSM	Shift Signaller Manager
SWAT	Subjective Workload Assessment Technique
TAM	Technical Acceptance Model
TOC	Train Operating Company
TOPS	Total Operations Processing System. A mainframe system used since 1970s.
TPWS	Train Protection & Warning System. Electronic loops designed to detect a SPAD and stop the train if one occurs.
TRTS	Train Ready To Start operated at a station to indicate train is ready to depart
TRUST	Train Running System TOPS monitors train movements for punctuality, used since 1980s
VDU	Visual Display Unit

Chapter 1: Introduction

1.1 Chapter overview

This chapter starts with a short introduction to the rail industry context and need for this research. The chapter then presents the research questions, and how each chapter contributes to answering these questions. The chapter ends with a brief introduction and clarification on key terminology that will be used throughout the thesis.

1.2 Rail context

Rail is a complex safety critical system in which the combination of staff, equipment and procedures provide safe operational performance. A system failure can put at risk the lives of staff, passengers, or the public. Automation technologies are being implemented to improve performance and capacity, whilst maintaining safety. Whilst these are designed to improve the system, they change the role of staff and their impact on staff is not fully understood. The industry seeks to improve the measurement of impact of these changes on staff and, more broadly, gain more value from data (RSSB 2017). The Rail Accident Investigation Board (RAIB, 2020) recommend improving workload assessments of signallers for this reason, particularly in their move to large, centralised workstations. Whilst this research is rail focused, the findings have implications staff across rail including drivers, signallers, control staff, and managers, and for other industries implementing automation technologies and changing staff roles.

1.3 Research perspective and scope

The research focuses on how personal physiological data can be used to measure human cognitive performance and workload. As automation technologies increase in rail, the balance between physical and cognitive elements of work is changing. Increased use of automation reduces the observability of workload: as staff physically move less, there is less workload to measure through observation. Yet there remains a need to understand and measure the workload of their cognitive monitoring and vigilance task. Alternative workload measures, such as self-assessed workload scales interrupt the task, or are completed after the task. An additional requirement, since COVID-19 restricted visitor access to signalling centres, is how to accurately measure workload remotely.

The research gap is finding a measure of human performance that can be applied in live operations to collect real-time data to measure the impact of technology changes, with minimal task interference. Such data could then inform management decisions. This research considers how temporal physiological data from wearable physiological measures could address this research gap. The temporal data from these devices provide a chronology of events in the task which, if patterns are detected, could indicate in future when staff are at risk of moving from good performance, into the higher risk areas of either underload or overload.

This research considers MWL and overall performance. In terms of MWL, previous theories and studies show humans are more likely to make errors when MWL is too high (overload) or too low (underload) (Reason, 1990). This implies that humans can perform more successfully when neither overloaded nor underloaded in terms of MWL. This research proposes to focus on the area between these two MWL levels, which is currently poorly defined. This range can be described by the Goldilocks Principle, representing a range that is 'just right', so called as it is analogous to the porridge not being too hot or too cold in the original children's story. In research around physical capacity at work, this Goldilocks Principle of productive work is work that stimulates a range of body systems whist allowing adequate rest and recovery (Holtermann et al., 2019). This research takes this principle and considers the implications for cognitive work. How could a range of cognitive work and effort be monitored that includes when it is 'just right', and thus sustainable. This scoping review considers whether physiological measures distinguish different levels of mental workload, from underload through productive work into overload.

In terms of overall performance, an overarching ethos of this research is to capture positive human performance and explore 'what does good look like', rather than focusing on when errors are made. This fits with the Safety II approach proposed by Hollnagel, where the aim is as much as possible goes right to achieve good performance in everyday work (Hollnagel, 2014). In rail, RSSB have developed questions to help derive performance indicators for the industry which includes not only 'what can go wrong?', but 'what does success look like?" (ORR, 2017). This research will consider whether physiological measures can detect factors associated with the task, or experience of workload, when performance is successful. If so, how could these contribute to our understanding of 'what does good look like'? Part of this includes avoiding some of the ironies of automation (Bainbridge, 1983) particularly around monitoring an ever increasingly automated system in rail signalling control.

To address ironies of automation regarding monitoring tasks, physiological measures could monitor alertness over time. This could show drops when an operator is bored due to low task load, or due to the vigilance decrement that is known to occur over time with a monitoring task (Mackworth, 1948). Secondly, regarding job satisfaction, when an operator needs to intervene at short notice or take over from automation, measures could show the effort, workload, or stress 'cost' on the part of the operator. Exploring these areas could build an understanding of what *is* sustainable in terms of cognitive effort.

The focus on railway signallers was a decision made in stages. The initial industry interview study identified a broad range of people who can impact rail operations including staff, passengers, and the public. In terms of industry priorities, train drivers and signallers would suit being the focus of the research, as they complete Safety Critical Tasks and are two of the biggest groups of staff. The scoping reviews of literature therefore considered driving and signalling tasks and roles. Proportionally less research was found that focused on signallers, presenting a viable research gap for the research. Secondly, to successfully apply physiological measures 'in the wild' it is important to control for the confounding variables such as physical movement and temperature. Taking all these factors into consideration, signallers were chosen as the focus of this research. These measures may, in future, also suit use in train cabs with drivers. A signalling centre, with temperature control and no vibration, was chosen as a best location at this exploratory stage of applying wearables measures in the rail industry.

1.4 Research aim

The research scope considers how personal data[1] could be used in rail to measure human cognitive performance. The research aim is to measure, in real-time, staff mental workload in live operations. The research focuses on cognitive performance, and the measurement of staff mental workload. This in turn would both improve our understanding of how staff achieve their cognitive tasks and inform future decision with respect to tasks and training in the rail industry. This is particularly pertinent to the implementation of automation technologies which are changing the roles of staff in the rail industry.

[1] This personal data angle fits with the 'My Life in Data' theme of the Horizon Centre for Doctoral Training (CDT), through which the PhD was completed. The Horizon CDT at University of Nottingham carries out interdisciplinary research within the Digital Economy theme, with a focus on digital identity, personal data, and data creativity.

1.5 Research questions

1. How can temporal physiological data from wearable measures contribute to MWL assessment in rail industry live operations?
2. What are the theoretical implications of individual physiological data to changes in MWL in a workplace setting?
3. What are staff perspectives on wearables and use of their personal physiological data?

1.6 Identifying a research gap

The initial remit of the research combined the rail industries sponsors request to research the impact of new technology on human performance in rail with the ‘My Life in Data’ theme of the PhD Horizon Centre for Doctoral Training Centre. A research gap was sought through a literature review and industry interviews. The interviews identified an industry challenge, so the research scope remained pertinent to industry. The interviews identified the challenge of detecting when staff are approaching their limits (underload or overload), including as new technology is introduced into rail. Drivers or signallers were identified as key roles to focus the research on. The literature review indicated signallers were the lesser studied group. Signallers were therefore selected as a focus for the research, after the industry interviews, with an understanding the findings from signallers could have implications for other staff roles in rail.

In addition to the studies presented here, a domain familiarisation activity was completed in signalling operations during an internship with Network Rail, a sponsor of the Horizon CDT PhD research programme. The internship occurred after the industry interviews, before the simulation study, helping to inform the remaining research. The internship was completed over three months at East Midlands Control Centre (EMCC) in Derby, a modern signalling centre with 10 workstations. Field observations were conducted of workstations, scenarios, teams, and individuals, to gain a realistic view of the complexity of the work environment (Roth and Patterson 2000) and contextual nature of expertise. Ethical approval was gained (see Appendix A) to observe, make field notes, and interview individual staff. All data was anonymised.

The naturalistic observation method is similar to ethnographic derived methods (such as Nardi 1997), providing the opportunity to be immersed in the daily lives of the people being studied. In the role of participant observer, the author acknowledges their presence will affect staff behaviour, though hoped to minimize such influences (Farrington-Darby and Wilson, 2006). As internship progressed, signallers and shift managers appeared to grow used to the researcher, known as habituation in field research (Robson 1993). The observations provided context for the literature on rail industry signalling operations, and the findings from the interview study around how human performances is assessed. The observations influenced the scope of the research by identifying the need for MWL measures that suit live operations. Specifically, signallers' workload was observed to vary greatly but peaks were difficult to predict as these tend to occur around incidents rather than only timetabled events. In addition, the workload in live operations is different than in a simulation, especially around communications (Sharples et al., 2011) such as phone calls. This presented a gap to research unobtrusive measures, to suit use in live operations, to detect MWL changes over time and the effort required. Measures would need to be used over extended periods of time compared to current measures to capture incidents as well as routine timetabled task demand.

1.7 Thesis structure

This thesis presents the progress made throughout this research. This research was conducted using a mixed methods approach, and in collaboration with industry. It includes two literature reviews and three studies.

Chapter 2 presents an overview of the rail industry context, and a scoping review of theories underpinning human cognitive performance, including information processing, the nature of expertise, and Mental Workload (MWL). Drivers and signallers are included to reflect research on staff in the rail industry. The additional detail provided on signallers' tasks reflects the subsequently decision to focus on applying wearables in signalling first, before driving. A range of current MWL measures are presented, including those tailored to the rail industry. Performance Shaping Factors (PSF) are mentioned with other broad themes peripheral to measuring human performance.

Chapter 3 presents the results from Study 1, an interview study with industry stakeholders. The study explored rail challenges that relate to human performance or new technology, individual attributes of performance, and how performance is assessed in rail and other transport industries. The study sought to identity who impacts rail operations, and industry priorities of who to focus the research on. This study informed the subsequent research focus of scope on railway signallers in live operations in later stages of the research.

Chapter 4 provides a scoping review of physiological measures and the critique used to select the measures suitable for signallers in live operations. The chapter includes, from the literature, the underlying physiology, and what is detected and can be inferred by the follow types of physiological measure: heart, skin, facial thermography; breathing; eye movement; electro-encephalography (EEG); and Functional near Infra-Red Spectroscopy (fNIRS). Based on this review, a decision was made to focus on HRV and EDA as measures of workload in rail signallers.

Chapter 5 presents the results from Study 2, a simulation study of twenty participants wearing physiological measures to infer MWL during a rail signalling task. Heart Rate Variability (HRV) and Electrodermal Activity (EDA) data were collected and compared to task demand and self-report workload. The study was conducted to provide an initial test of methods, equipment, data processing and data visualisation prior to future live trialling of measures with industry. Results showed what aspects of workload different measures were sensitive to, with storyboards graphing dynamic changes over time to show the complexity of relationships between HRV, EDA, task demand and subjective workload.

Chapter 6 presents the results from Study 3 the perspectives of staff to the use of wearable physiological measures. The interview study explored signalling staff perspectives on the potential use of wearable measures in the workplace in future. The study method combined semi-structured interviews and surveys with rating scales. Analysis considered to what extent personal attitude to change could predict technology acceptance. The study found wearable devices suit use in the live operational environment, with the wrist strap rated the most suitable due to low distraction and perceived ease of use. In terms of data use, themes included perceived usefulness, anonymity, and trust.

Chapter 7 presents answers to the research questions and novel contributions of this research. The discussion draws together findings from the industry interviews, simulation study, and perspectives and attitude studies. The discussion includes how physiological measures can contribute to MWL assessment, staff perspectives and attitudes on their use, theoretical implications, and implications for industry.

Chapter 8 completes the thesis by presenting the contributions and conclusions of the research, and recommendations for future work on the measurement of human performance and the development of physiological measures.

1.8 COVID-19 impact statement

The lockdowns put in place in response to COVID-19 had a direct impact on this research. A live trial using wearable physiological measures was planned in full but cancelled on 13th March 2020, the week before it was due to run at East Midlands Control Centre (EMCC), Derby. Twelve signallers, working at EMCC, were due to wear a wrist strap and chest strap for 4 to 5 hours, and respond to a self-assessment workload measure, whilst completing half their shift at a live operational workstation. The study aimed to evaluate whether wearable physiological data could:

- Indicate the mental effort of signalling staff to complete their task.
- Match the results of existing workload measures (Instantaneous Workload Scale).
- Detect task related events that contribute to task load.

Planning for this trial began with Network Rail in March 2019, with the RMT Union informed of the study in April 2019, and University ethics approval gained in February 2020. On Friday 13th March 2020 all operational site access was closed to visitors to minimise the risk to operational staff of COVID-19. Restrictions to visitor access remain in place to date.

An alternative research approach was needed to replace the planned live trial of wearables. This research was subsequently adapted, to ensure the research could go ahead irrespective of changes in COVID-19 restrictions. Regular contact was maintained with Network Rail. By May 2020 a remote interview study with operational staff was agreed with supervisors, RSSB and Network Rail. This change ensured the study would be more likely to be able to proceed if visitor access remained restricted. University Ethics was approved in September 2020. Interviews began in November 2020. Interviews were subsequently halted for 5 weeks following staff cases of COVID-19 and staff self-isolating. Interviews were successfully completed in January 2021. The findings from this staff perspective study forms part of this thesis.

1.9 Terminology

There are terms used throughout this research that benefit from introduction at this stage: Cognitive vs mental workload; Human Performance vs competency; and Psychophysiology vs physiological.

The term mental relates to the mind and is used interchangeably with cognitive to refer to the processes of thinking (Cambridge Dictionary, 2021) and to distinguish from the physical. MWL is a construct that encapsulates task demand, how individuals experience workload and performance (Sharples, 2019). Cognitive ergonomics includes MWL, information processing, reaction, decision making, and stress (Wilson and Sharples, 2015). In this research the term mental will be used throughout.

Human performance in this research refers to the capacity of any human to achieve a task. Performance can describe the act of achieving a task and a measurable outcome. This deserves clarification as, in rail, performance refers to operational performance and human performance is thought of as 'competency' (Fowler et al., 2019). In this thesis human performance is used as a broad term that refers to achieving a task. This performance is influenced by both competency and MWL, with successful performance outcomes reliant on sufficient competency, and a level of MWL that is between the states of underload and overload. Further description of MWL is provided in Chapter 2 Theories.

Physiological measures detect a physical aspect of bodily function. The term physiological measure encapsulates both the device and data collected, and wearable measure refers to the device that fit on the body. When physiological measures are applied to cognitive activity they can be referred to as psychophysiological measures. In this research the term 'physiological' will refer to the data that the measures detect and psychophysiology as a field of research.

Alertness levels are relevant to this research, with a focus on cognitive 'alertness'. Various terms are used in academic literature and in industry that refer to a similar construct. Vigilance refers to the ability to maintain focus of attention over a period of time (Davies and Parasuraman, 1982). In psychophysiological literature the term arousal is used (Hugdahl, 1995) to describe a level of cortical, behavioural, or autonomic activity. The common usage of arousal can imply sexual arousal. Tonic alertness can be viewed as synonymous with vigilance and sustained attention (Oken et al., 2006). In the rail industry concentration is the term recognised as equivalent to vigilance (Pickup et al., 2010). In this thesis the term 'alertness' will be used throughout.

Here the term rail and railway are used interchangeably to refer to the rail industry. The rail industry encompasses all operations of railway technology which transport goods and passengers from one place to another along railway tracks.

Chapter 2: Industry context and theories of human performance

2.1 Chapter overview

This chapter presents an overview of the rail industry context, and a scoping review of theories underpinning human cognitive performance, including information processing, the nature of expertise, and Mental Workload (MWL). Drivers and signallers are included to reflect research on staff in the rail industry. The additional detail provided on signallers' tasks reflects the subsequently decision to focus on applying wearables in signalling first, before driving. A range of current MWL measures are presented, including those tailored to the rail industry. Performance Shaping Factors (PSF) are mentioned with other broad themes peripheral to measuring human performance.

2.2 Introduction

This research considers how changes in procedures and technologies impact human performance, and how to assess the impact on the cognitive aspects of their workload. This aims to inform choices made around changes implemented in future. This chapter presents the context of the rail industry, relevant theories of cognition and mental workload measures identified in a scoping review. In doing so, this chapter contributes to the first two research questions by identifying existing MWL theories and measures and describing rail operations.

This research considers human performance from the perspective of human cognitive performance (see shaded area in Figure 2-1), in the context of external changes to tasks being performed. Within this, the research considers both the Safety I view of 'what goes wrong' and the Safety II view of 'what goes right'. This is to cover the full range of performance in live operations.

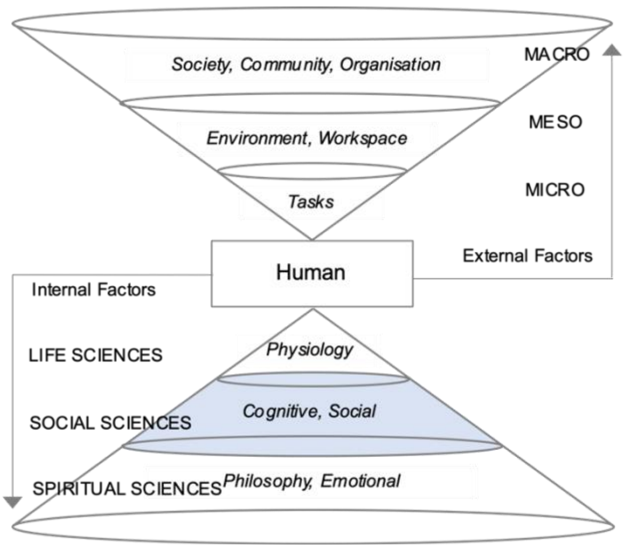


Figure 2-1 Qualitative Methodology in Ergonomics, Hignett (2001, p.210).

Shaded area shows focus of scoping review

Safety I and Safety II are contrasting approaches to safety, of relevance to this research. Safety I is an established approach to safety that seeks to reduce danger, risk, and injury (Vaughan, 1997). It does this through investigating incidents and accidents, seeking to identify causes, implementing preventative measures, and monitoring lagging after the event e.g. reduced injury or incident rates (Lingard et al., 2013). In Safety I, humans can be viewed as a liability (Hollnagel et al., 2013) and cause of safety failures. Safety II, in comparison, proposes a different perspective on Safety, concerned with ensuring 'as many things as possible go right' (Hollnagel, 2017). It seeks to be a proactive, adaptive approach to safety, by monitoring success. In Safety II, humans are seen as a resource for system flexibility and resilience (Hollnagel et al., 2013). Safety II is in its infancy, along with other new approaches to safety termed New View including Resilience Engineering, Human and Organisational Performance (HOP) (Conklin, 2012), and Safety Differently (Dekker, 2014). New View and Safety I converge on the concept that safety involves the management of risk (Bergström et al., 2015). In other features, the two diverge. Critics of Safety II note there are no published, peer reviewed, empirical evidence that Safety II improves safety (Cooper, 2020), and it lacks measures. Cooper suggests, in their position paper, that New View should not replace a Safety I approach, as New View offers a perspective only rather than a methodology (Cooper, 2020). Cooper postulates that cultural differences could explain the differences between Safety I and II, being at different ends of a cultural continuum between cultural tightness and looseness, drawing on Gelfand's assessment of different cultural responses to COVID (Gelfand et al., 2021). Some situations require a tight, strong Safety I cultural approach to manage safety, some a cultural looseness of Safety II as appropriate COVID (Gelfand et al., 2021). The choice between Safety I and Safety II may not need to be an either or, but more for what proportion of each is beneficial.

In the rail industry, maintaining a high standard of safety remains an important goal. The railways continue to benefit from a Safety I approach to improve safety following accidents. This includes the work by accident investigation organisations across the world such as Rail Accident Investigation Branch (RAIB) in the UK, National Transport Safety Board (NTSB) America, Australian Transport Safety Bureau (ATSB), European Railway Agency, and Japan Transport Safety Board. More recently in rail there has been interest in applying Safety II type principle such as considering "What does success look like?" to understand fatigue (ORR 2017). Where human performance research in rail has previously focused on what goes wrong, a more positive approach may be the monitoring of what goes right every day and system strengths (McDonald, 2021) and identifying protective factors in near misses (Thoroman et al., 2019). One reason for doing this is, as incremental safety improvements work, the occasions to measure major safety failures reduce. In Britain 4,917[2] days occurred without a fatality involving a passenger train between the incident at Grayrigg in 2007 and the Stonehaven derailment in 2020. Recent research in rail does appear to have shifted away from failure analysis towards human activities that *maintain*

[2] The accident at Grayrigg, 23 February 2007, caused 1 death of 109 on board. Number is intervening days.

safety (Ryan et al., 2021). As part of that endeavour, this research aims to research how the MWL of staff can be monitored whilst achieving high levels of safety and consider how such data could provide feedback through live status to managers. This research considers how to assess workload changes over time, to detect variation over time, including when performance outcomes are acceptable. In broad terms, therefore, this research has a Safety II perspective.

This research considers the *range* of individuals' performance, of effective and ineffective workload (Xie and Salvendy, 2000), including the Safety II premise of what can be learnt about everyday success and Work-As-Done (Hollnagel, 2014) rather than focusing solely on human errors (a more Safety I approach). The justification for this has two parts: monitoring the range of human effort that underlies daily successful operational performance could improve industry understanding of the impact of change; and detecting patterns in data leading up to incidents to provide leading indicators of deterioration in human performance sufficient to warrant operational intervention (e.g. a break, or allocation of additional staff resource). These fit a Safety II approach seeks to understand and measure what goes *right* to result in intended and acceptable outcomes (Hollnagel, 2014).

The remaining sections of this chapter provide an overview of the outcome of a scoping review of the theories on cognition and measurement relevant to front line rail staff performance. It starts by introducing the rail industry context and safety critical tasks. It then presents theories of information processing, expertise, and MWL, highlighting those applicable to complex continuous tasks in a dynamic field setting. It goes on to the MWL measures used in rail.

2.3 Rail industry context

2.3.1 Rail as a sociotechnical system

The rail industry is a large complex socio-technical system (Wilson, 2014), whose operations rely on skilled human interaction with physical engineering such as trains, stations, and signals. Rasmussen's socio-technical system model of risk management (Rasmussen, 1997) represents the hierarchy of control in such a system (Figure 2-2).

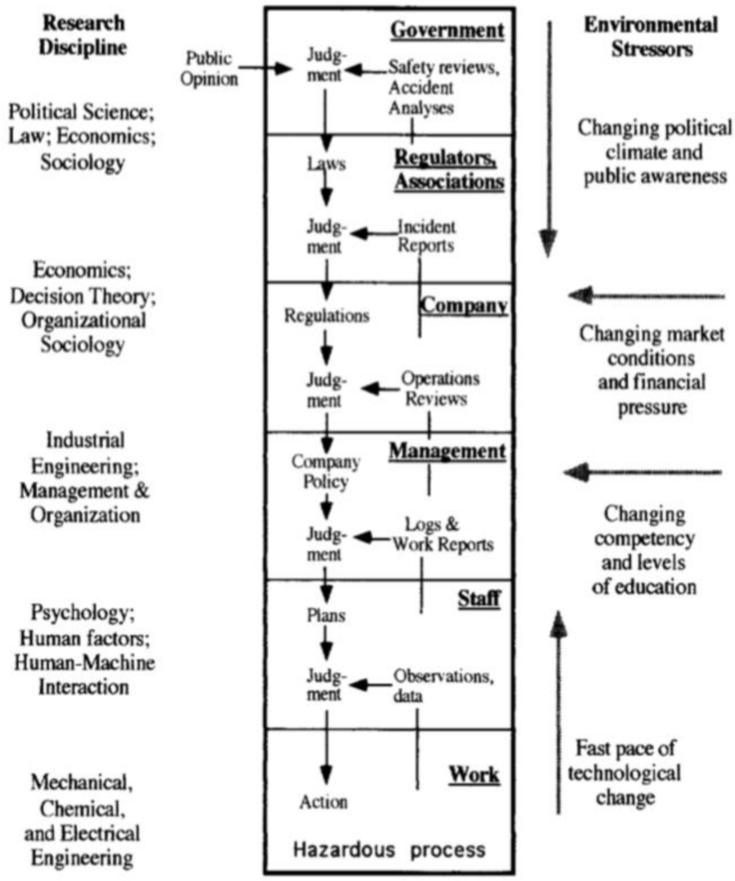


Figure 2-2 The socio-technical system involved in risk management (Rasmussen 1997)

Rasmussen notes two challenges faced in the present dynamic society: the very fast pace of change of technology at the operational level at odds with ever slower responses to change at every level up the hierarchy; and the ever-increasing integration of systems making modelling in isolation increasingly difficult. The areas of the hierarchy in Figure 2-2 of focus in this research are the human factors discipline and staff process level, with implications for management and work processes in relation to changing technology and competency. The research considers how staff MWL data could provide feedback across levels of the system as a form of technology aided mutual monitoring. Within a team at work, mutual monitoring means colleagues intervene and assist with workload in a timely manner e.g. by answering a phone work live operations. This informal team support is not visible to those outside the control room. Measures that could visualise changing MWL could increase the visibility to managers outside of the control room. This could, in turn, inform managers' decisions for front line staff to support a systems approach to accident prevention through vertical integration (Thoroman et al., 2019).

Various automation technologies are being introduced in rail to assist drivers and signallers. The intention is to increase capacity, whilst maintaining safety. The challenge is how to provide data for rail managers on the impact of these changes, to inform their future judgements, to ensure the intended benefits of new technologies are realised. One challenge specific to cognitive performance is determining whether changes increase the risk of staff overload or underload (deWaard 1996). Underload, or excessively low mental demands are detrimental to performance (Young and Stanton, 2002). A research gap exists in the collection of real-time data on human performance. If collected, such data collected at the staff and work levels of Figure 2-2 would be used to seek patterns, and ultimately predict, impact of changing technologies and inform management decisions at the management and company levels of Figure 2-2.

2.3.2 Drivers and signaller tasks and attributes

Safety Critical Tasks are those tasks that involve responsibility for the safety of passengers, staff, or the public (ORR 2017). Staff performing their tasks successfully are a barrier function (Golightly et al., 2013) or 'last line of defence' (RAIB 2020) to safety. Drivers and signallers perform safety critical tasks, and both experience the impact of new technologies in the train cab and signalling control centres. The research considered both drivers and signallers in the first Industry Interviews study, before the focus narrowed to signallers in the Simulation and Staff Attitudes and Perspectives studies. This scoping review includes our current theoretical understanding of the cognitive tasks, and workload, of drivers and signallers as experts.

Certain individual attributes help drivers and signallers to achieve their tasks and handle the complexity and responsibility of their roles. These individual attributes such as attention and communication are addressed initially during recruitment. Driver recruitment includes tests of attention, concentration/vigilance, memory, and communication (Train Driver 2021). Signaller recruitment includes tests of situational judgement, and numerical, verbal reasoning, and inductive reasoning (Practice4Me 2021). Once recruited, drivers and signallers become experts through extensive training, followed by supervised live experience, before being passed as competent.

It should be noted that across world the role of signaller varies, with alternative terms of dispatcher and controller. In Australia the controller is more strategic, and the signaller is more a tactical role. Train controllers are responsible for managing the strategic overview of the whole rail network including recovery from disruption (Dorrian et al., 2011). Signallers cover a more tactical role turning controller's plans into actions by operating the points to set routes (Dorrian et al., 2011). In the USA, railroad dispatchers are responsible for the safe, efficient movement of trains and the protection of the workforce working on the track (Gertler and Vaile 2007). In this regard they are responsible for traffic management. In the USA signallers operate the points, like signallers in the USA. In Britain, dispatchers are station platform staff who communicate the train is ready to depart. The signaller plans, sets routes, and authorises trains to move through the rail infrastructure while ensuring separation between all trains. They make decisions on the order of trains through junctions for effective traffic management, particularly during disruption. (Balfe et al., 2015). In Britain, minor late running is all handled by the signaller. The controllers in Britain handle major diversions (Farrington-Darby et al., 2006), coordinating between train operating companies and signallers. The implications of the findings of this research are applicable to dispatchers and controllers around the world who monitor, plan, and must recover from operational disruption. In this thesis the term signaller is used throughout.

Rail is a dynamic work setting, where keeping trains running safely and on-time are balanced continuously. This forms a paradox as keeping time *and* driving safely can conflict (Naweed and Aitken 2014). External task load factors, affecting drivers and signallers, include regular imposed time constraints (e.g., station dwell times, or permitting access to track by maintenance staff between trains running). Both drivers and signallers have the authority to stop a train to maintain safety, however, normal running relies on trains moving at line speed to keep to the timetable. Any disruption, or perturbation, affects operational performance and can add complexity and uncertainty to the roles of both staff. Subsequent commercial pressures come from the costs incurred through Delay Attribution (DAB) (Network Rail, 2023). Delay Attribution is used in rail to determine the cause of any delays and whether the train operator or infrastructure provider are liable for the cost incurred of any delays.

Whilst drivers and signallers perform safety critical tasks with time constraints, there are differences in their roles that affect the proportion and types of cognition tasks required. Drivers are responsible for the safe operation of the train, including maintaining appropriate speed, stopping at stations, warning other rail users (using the horn) and communicating with signallers and passengers if required (e.g., during delays) (Ryan et al., 2021). Drivers are responsible for the safety of all onboard, themselves included. In terms of cognitive skills, drivers require good memory (including system knowledge), monitoring skills, sustained attention, and decision making (Buksh et al., 2013, Naweed 2014). Their 'route knowledge' allows drivers to anticipate, which is key to their role (Buksh et al., 2013, Balfe et al., 2017). Signallers are responsible for monitoring and authorising train movements, anticipating delays and poor traffic flow, implementing speed restrictions, and communicating with track workers, control, station staff and drivers, and may also operate level crossings (Golightly and Young, 2022; Ryan et al., 2021; Sharples et al., 2011). Signallers are present in time, but in comparison to drivers, they are remote from the track, controlling signals and points from a box, or signalling centre. Signaller experiential knowledge is important, built through experience and comprising local geographical knowledge and likely patterns of trains (service patterns) to provide safe and efficient performance (Golightly and Young 2022, RAIB 2020). Compared to drivers their role more often requires coordinating with other staff to avoid or address delays. Signallers also require good skills in communication, planning and prioritising, problem solving and decision making, collaborating and being vigilant and resilient under pressure (PENNA 2018). These Non-Technical Skills (NTS) are applicable to railways around the world (Flin et al., 2017; Jarosz et al., 2021; Madigan et al., 2015; Nayak et al., 2018).

2.3.3 The impact of new technology on driver and signaller tasks

Over its history, the railway in Britain has benefited from new technologies. The railway itself was a new technology and disruptive innovation in 1825. Initially constructed for freight, the public began travelling in 1830. The railway then grew, with construction of new railways peaking in 1847 (Bradley 2016). The bells allowing signallers to communicate are still in use in over 500 lever frame signal boxes today. Each innovation adds to the mixed ages of legacy technology that must work in parallel. On the trains, by 1900, bogies (part of the suspension) provided a smoother ride (Bluebell Railway, 2021). Steam locomotives, then diesel and electric, allowed trains to run faster, and greener. Today Britain has 15,904km of route, of which 38% is electrified

(ORR 2020). Added safety technologies, such as Train Protection and Warning System (TPWS), have reduced the risk of collisions and led to British railways being amongst the safest in Europe (ORR 2013). Drivers, until the last decade, had to get out of the train cab to speak to signallers via a signal post telephone. Communications in the cab have since become possible using Global System for Mobile Communications-Railway (GSM-R), installed between 2009 - 2016 (Wikipedia 2021). The 7-year implementation indicates how long 'new' technologies take to integrate into an industry, like the railway, that must continue to operate daily. So, in train cabs, the 'telephone' is a relatively new technology.

In Britain trains currently operate at varying levels of automation. Mainline trains are controlled manually by a driver, with safety systems such as TPWS that applies the brakes if the train passes a signal at danger without authority, or approaches buffers or a signal at danger too fast. At the high end of automation is Docklands Light Railway (DLR) in London. The DLR is an automated light metro system that is driverless (see passenger front seat view in Figure 2-3). Train movements are monitored from a control room. A staff member is onboard to assist passengers. Thameslink, use Automatic Train Operation (ATO) in the 'core area' of London, with drivers taking back manual control outside of this area. Driver cognition is crucial during transitions between different forms of driving, or mixed mode driving (Buksh et al., 2013).

Figure 2-3 Driverless DLR passenger front view (Authors image, 21.11.2017)

Signalling operations in Britain are undergoing a long-term plan to consolidate 800 signal boxes into 12 centralised Rail Operating Centres (ROCs)

(Network Rail 2019). These centres benefit from upgrades in signalling technology to enable one signaller to control a larger area of route. This increase in area is evident in Figure 2-4. The figure shows the small route area controlled from one lever frame signal box at Egginton[3] compared with workstations. Determining the area controlled by a workstation is one way to balance the task demand of signallers. At Derby East Midlands Control Centre (EMCC), the areas covered by the Nottingham workstation is smaller than others as it routes a high number of trains and contains a busy station with multiple platforms. Netherfield covers a larger area as it routes fewer trains, with no major station. The challenge is their area is difficult to alter once in place.

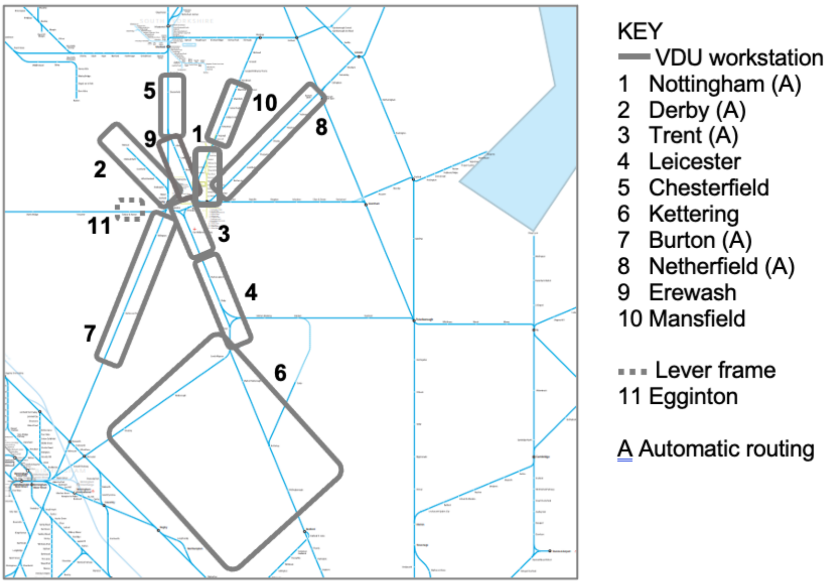


Figure 2-4 Area controlled by VDU workstation at Derby, versus a Lever Frame signal box

Presented in rank order of number of timetabled trains including freight.

The advantage of the move to a modern centre is the opportunity to add assistive technology such as Automatic Route Setting (ARS) to signalling control. Half the workstations at Derby have a form of ARS (see Figure 2-4 key). ARS makes and implements decisions autonomously (Balfe et al., 2015) unless the human overrides. Compared to other industry automation (Vagia et al., 2016), this is a high-level of automation. The ARS makes decisions

[3] Egginton opened in 1800s and remains in operation today (Derby Signalling 2021).

based on timetabled information. A simulator study found ARS reduced mental workload and improved performance (Balfe et al., 2015), particularly during normal running. Normal running in this case refers to trains running to timetable, without delays. In Britain, per year, approximately 10% of passenger trains are more than 10 minutes late at their destination (ORR 2022). Dealing with delay is therefore a regular occurrence in Britain, although this varies across operators and regions (ORR 2022). Frequency of delays vary around the world based on factors from the condition of trains and track in Sweden (Palmqvist et al., 2017), to overcrowding in Japan (Fujiyama 2018). If trains are running late, signallers in Britain need to check which train ARS plans to route first to determine whether to alter that plan. This can increase their workload if they must check each train on repeated occasions. The Balfe study noted the workload reduction is not as large during disrupted running, and more complex effects of disruption may not have appeared in the simulation study (Balfe et al., 2015). In live operations, once ARS sets a route the signaller can see it, procedures, recommend they do not change the choice if there is an approaching train. In addition, the British control system, the signaller can only set routes for a train when that train will be the first to reach a junction. This means any intervention must be carefully timed. It cannot be corrected in future as is possible in for Train Control Officers in the Netherlands, or Service Direction Leaders in Germany. More recently the Rail Union have raised concerns that ARS can increase signaller workload and have requested clarification on the allocation of function between the signaller and the automation with regulating (deciding which train goes first) (RMT 2020). These staff concerns suggest the reality of automation in the full range of live operational scenarios, is more complex than a blanket reduction in MWL indicated by the simulation study. This adds to the case for measures that can capture MWL during live operations, including during disruption.

The move to signalling centres is intended to benefit the signallers in other ways. Firstly, a positive improvement is the manual handling of heavy levers in a lever frame (left picture in Figure 2-5) has been replaced with buttons and tracker ball controls (right picture in Figure 2-5). This move has, however, introduced different challenges. It changes the balance between physical and cognitive elements of work, with increasingly cognitive rather than physical task components (Sharples and Megaw 2015). The disadvantages of this change are it includes an increased risk of underload, and associated difficulty of remaining vigilant whilst completing a monitoring task. In train driving the risk of reduced vigilance that comes from fatigue or underload have been address by the adoption of in-cab vigilance devices (that apply the brakes if the driver fails to respond in time). The implications of this are that it remains important in rail to measure MWL (both underload and overload), but the industry increasingly requires a measure that is less reliant on observable visual information.

A disadvantage of larger centralised control is the loss of local knowledge. Control becomes remote from location. Lever frame boxes have a view of the track they control (e.g., refer to Egginton in Figure 2-5) supplemented by the sounds, smells and vibrations on location, of weather conditions and the state

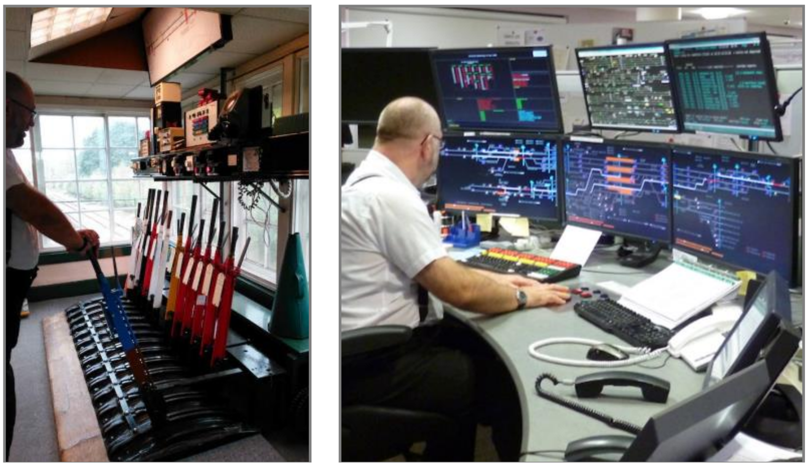


Figure 2-5 Comparing current signalling technologies

Egginton Lever frame signal box (left) [Author 03.08.2018] Derby VDU workstation (right) [Author 12.12.2018] of trains and the track. Signalling in centres is remote from the location it controls, relying on visual information on screens and audible alarms. In addition, the areas become too great for signallers to have a high level of local knowledge. They may not know or live near the area. Automation can route trains when they are running normally, to timetable. When signallers deal with situations or incident at stations, level crossings or any degraded working their local knowledge can assist them.

The next step change in British rail operations is the European Rail Traffic Management System (ERTMS) which is being introduced over the next 30 years (ORR 2021, ERTMS 2021). Signalling will be displayed in the train cab, replacing line side signals, and allow an increase in capacity by spacing trains relative to each other rather than to line side signals. Signallers would have more trains in their area of control, with the automation designed to route timetabled trains. Signallers would need to intervene when times or patterns of events do not match the timetable.

In summary, new technologies impact both drivers and signallers. New technologies take time to fully implement, meaning a telephone in the train cab is a relatively new technology and requiring staff to work with mixed ages of technology. Control is moving remote with an increase in automation. The automation can reduce staff workload during normal running but requires staff to intervene when there are delays or incidents.

Having presented the rail context of this research, with examples of the impact of new technology, the following section will present the theory underpinning our current understanding of human cognition and expertise.

2.4 Understanding the cognitive performance of experts

Theories of cognition and expertise, combined with neurology help explain human performance and why certain cognitive tasks, or combinations of tasks, are easy or hard for humans to complete. This section of the scoping review highlights current theories and concepts, models of MWL, and how current measures are used (Colquhoun et al., 2014). The theories and concept to understand the cognitive performance of experts include information processing, alertness, fatigue, expertise, decision making and the skills rules knowledge framework.

2.4.1 Information processing

The term ‘cognitive’ refers to “the action or process of knowing” (Oxford English Dictionary 2021). This happens in a sequence of stages summarised in Figure 2-6, adapted from Wickens (1999) and simplified from Sharples and Megaw (2015). This model has been developed and adapted over time to reflect the changing theoretical understanding of information processing, with the term *attentional resources* replacing *limited processing capacity*. Initially sensory information is perceived by our senses[4] and processed in the cortex, initially separately for each sense. Perception then draws on this processed information, combined with memory, for a response to be selected and executed. Attentional resources acknowledges that humans have limits on their perception, memory and response execution (Sharples and Megaw 2015) that constrain how they can complete a task. The sensory information processing combined with long-term memory could be automatic (involuntary), with no conscious awareness, or voluntary with varying degrees of conscious awareness. It is important to note, when considering MWL measures, the large number of processes are not conscious so cannot be subjectively assessed by asking people to report their workload (Meijan and Mulder 1992).

[4] Railway staff must meet requirements for hearing and sight e.g. no colour blindness.

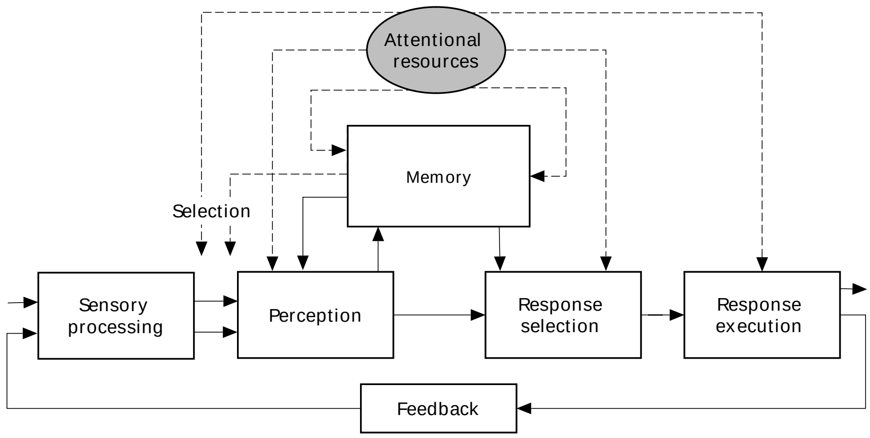


Figure 2-6 A general model of human information processing (Wickens et al., 1999)

Wickens' Multiple Resources Theory (2002) considers demand, and the impact of resource overlap and allocation policy, to understand the extent to which tasks can be completed in parallel. A study that applied Multiple Resources Theory to the work of signallers in rail found that aspects of the signallers' task required the same types of resource, so overload was a risk despite newer signalling technology (Krehl and Balfe, 2014). The observational data from the study identified twelve typical signaller tasks. The most frequent were monitoring, setting routes, communication (with colleagues and supervisors), and referring to information in the timetable or Train Running System[5] (TRUST). Signallers use both timetable (the planned schedule) and TRUST information (current train punctuality data) to plan which order to route trains. This enables signallers to balance maintaining the timetable and, when possible, recover from previous delays. The observer noted that during more than half the time of observing tasks were completed in combination such as whilst maintaining monitoring or setting routes. Using the newer VDU based system (a set of computer screens as shown in the right image in Figure 2-5), compared to an older NX Panel (an abbreviation of eNtrance-eXit control panel, with physical push/pull buttons for route setting), was found to increase referencing to timetable or TRUST

[5] Train Running System (TRUST) is a system contains train operational data as compared to schedule. It monitors train movements for punctuality and has been used since the 1980s.

Information, and increased attending to audio signals. In comparing these findings using the Multiple Resources Theory found that signaller tasks, such as monitoring and routing, make use of similar cognitive resources. Whilst VDU workstations can cover larger areas with the use of automation, the interaction with the control system is largely the same. The study concluded that overload remained a danger, particularly during non-normal working conditions.

2.4.2 Vigilance and alertness

Vigilance is the ability to maintain focus of attention over prolonged periods of time (Davies and Parasuraman, 1982). Vigilance as a construct is pertinent to drivers, signallers, and to this research when considering the unobservable aspects of MWL. The subject of vigilance has also become more pertinent as automation has increased (Parasuraman and Riley, 1997), changing human roles to more of a monitoring task. Vigilance is considered here, along with the associated constructs of attention, alertness, arousal, and concentration, and how they relate to MWL.

The derivation of vigilance is the Latin for ‘keeping awake’ and alert. It is a term used by cognitive neuroscience and psychology researchers (Oken et al., 2006). In psychology research, the term vigilance relates to signal detection and readiness to respond. Vigilance decrement describes a decline in attention performance, that emerged from research with radar and sonar operators during World War II (Mackworth, 1948). The research found vigilance wanes quickly in a monitoring task, with a 15% decline in signal detection after 30 minutes. It has since been found this vigilance decrement affects experienced and naïve watchkeepers alike (Warm et al., 2008), and has been demonstrated in operational and laboratory settings (Baker, 1962; Pigeau et al., 1995). In rail, train drivers must be vigilant (Gillis, 2016), with a test for vigilance forming part of selection for drivers (Train Driver, 2021). The importance of vigilance in railway operations is evidenced by the presence of a vigilance device in train cabs, alarm if the driver has not responded to any controls for a period of time (Whitlock et al., 2018).

Attention, in comparison, derives from the Latin ‘to take notice of’. Attention is more focused on cortex activation that supports information processing (Mesulam, 1990). Where vigilance and attention overlap are in the term sustained attention, which describes the ability to attend over time (Davies and Parasuraman, 1982). Vigilance can be viewed, therefore, as synonymous with sustained attention (Parasuraman et al., 1998). In rail, errors in attention of train drivers were identified as a contributing factor to certain incidents, in particular Signal Passed At Dangers (SPADs) and Train Protection and Warning System (TPWS) activation, (Madigan et al., 2016), where a train has passed a stop signal and the train breaks automatically applied to protect trains ahead. The term alertness refers to the state or quality of being alert, whilst also acknowledging the part played by cognitive processing. There are two types of alertness: phasic, which focuses on the orienting response to stimuli (Sokolov, 1963); and tonic alertness, which can be viewed as synonymous with vigilance and sustained attention (Oken et al., 2006). This way of viewing vigilance being tonic alertness and sustained attention, implies a degree of arousal on the sleep wake axis and the level of cognitive performance (Oken et al., 2006). This view suits this research, in that it considers both the physiological underlying state, and the impact this has on cognitive performance. When vigilance is good, tonic alertness is sufficient for cognitive performance to be sustained over time. Oken recommends future studies analysis the finer temporal aspects of this physiologic-performance relationships.

Arousal, mentioned briefly above, overlaps with alertness. Arousal is conceptually distinct, referring to a state associated with the level of activation of the cortex in the sleep-wake cycle (Oken et al., 2006). Clinical neurophysiologist may use the term vigilance level here to refer to arousal level on the sleep-wake spectrum, without reference to cognition (Oken et al., 2006). In psychophysiological literature the term arousal describes a level of cortical, behavioural, or autonomic activity (Hugdahl 1995). The common usage of arousal can imply sexual arousal. This is not the focus here, where instead the focus is on cognitive ‘alertness’.

Concentration is a construct that includes attention and focus on one thing, but not explicitly specifying a duration. As a term it is relevant here as it is recognised by SMEs in rail and is mentioned as a Non-Technical Skill in Rail (see Table 3-2), and during the development of ODEC as a workload measure for railway signallers, it was a term found to be relevant across all four elements considered relevant to MWL namely operational infrastructure, indicators, processes and service pattern (Pickup et al., 2010).

Regarding how vigilance relates to MWL and task demand, traditionally vigilance has been viewed as ‘benign’, with monotonous tasks assumed to be mentally undemanding (Warm et al., 2008). More recently, however the opposite has been indicated suggesting that, when task demand is low, greater effort is required for monitoring tasks thus increasing reported MWL (Warm et al., 2008). Warm *et al.* found undemanding tasks gained surprisingly high MWL rating on NASA TLX, especially effort and frustration (Warm et al., 1996). The researchers proposed this could reflect an increase in attentional resource exerted to overcome the tedium of the task. In contrast, when task demand is high, high vigilance is linked to increased MWL and stress (Baker, 1962).

In summary, vigilance overlaps with several associated constructs are relevant to this research as, with an increasingly monitoring task, they are associated with unobservable mental workload, and consider the temporal interactions between physiology and performance. The focus here is on cognitive ‘alertness’, or tonic alertness, which can be viewed as synonymous with vigilance and sustained attention (Oken et al., 2006). In this research the term ‘alertness’ will be used throughout to refer to the construct of vigilance.

2.4.3 Fatigue

Fatigue is a significant performance shaping factor in railway operations (Kyriakidis et al., 2015). Yet, like MWL, the search for measures of fatigue continues and the management of fatigue faces organisational barriers in the rail industry. Consideration of how fatigue is measured and managed in rail could help inform the introduce of new physiological data into rail.

Fatigue can occur due to both sleep related and task related factors (May and Baldwin, 2009). The UK rail regulator defines fatigue as 'a state of perceived weariness that can result from prolonged working, heavy workload, insufficient rest and inadequate sleep' (ORR 2012). Research on sleep related factors shows higher fatigue levels are associated with less observed sleep or lack of sleep (Caldwell et al., 2009; Darwent et al., 2015; Dawson and McCulloch, 2005; Dorrian et al., 2011; Young and Steel, 2017), with a clear relationship between the length of time off, time-of-day and the amount of sleep obtained (Roach et al., 2003). In addition, circadian disruptions (Caldwell et al., 2009) or sleep pattern interruption (Åkerstedt 1991) can contribute to fatigue. Task related factors associated with fatigue include long duty periods (Caldwell et al., 2009) such as 9 -12 hour shifts (Filtness and Naweed, 2017), monotonous tasks with low task demand (Anund et al., 2015), or tasks requiring high mental effort and sustained vigilance (Phillips et al., 2010) which does not allow for recovery (Dunn and Williamson, 2012). The effort and vigilance (alertness) element of this are where physiological data may detect changes.

Fatigue and sleepiness are common in transport operations and a significant cause of safety-critical events (Anund et al., 2015). To date there is proportionally more research in road and aviation industries than in rail (Anund et al., 2015), such as aviation (Caldwell et al., 2009) and automotive (May and Baldmin 2009). In the rail industry much fatigue research has focused on train drivers (Sussman and Coplen 2000), identifying fatigue as affecting train drivers (Filtness and Naweed, 2017) especially freight drivers, and especially at night. Fatigue of train drivers can affect passenger services too, as demonstrated by the accident of a train collision with buffers stops at a terminus station after the driver had a micro sleep at the end of night shift after 18 hrs awake (RAIB (2017). Other research across the world confirms fatigue can negatively affect signallers and dispatchers (Sussman and Coplen 2000, Dorrian, Baulk and Dawson 2011, RAIB 2020) and other shift workers (Dorrian, Baulk and Dawson 2011).

Current methods to manage fatigue include employer and employee controls. Employer controls include: limits on working hours across transport industries (Jones et al., 2005) and in rail (Young and Steel 2017); rosters (Ashton and Fowler 2005); fatigue modelling, including biomathematical, to predict fatigue risk prior to implementing a roster (Darwent et al., 2015; Filtness and Naweed, 2017; Young and Steel, 2017); and fatigue monitoring such as in USA, Canada and Australia (ONRSR 2021, Transport Canada 2022). Employee controls include self-regulation and sleep management, and techniques to counter fatigue such as going for a walk, having a nap (Filtness and Naweed, 2017), and consuming caffeine (Dunn and Williamson, 2012, Anund et al., 2015). Individuals can, however, be bad at judging their own fatigued state (Martindale, 2012).

Despite the countermeasures mentioned above, research in the UK and Australian rail industry identified that organisational culture can be a barrier to fatigue management. These may be relevant to the introduction of physiological measures. There appears to be an ingrained culture surrounding fatigue (Young and Steel 2017), in a highly reactive organisational culture where fear has developed, so even the mention of fatigue is considered taboo (Filtness and Naweed, 2017). This is potentially due to peer pressure or knowing it will increase the workload of colleagues (Young and Steel 2017), or motivated by the extra income, despite knowing the risks (Filtness and Naweed, 2017). It is thought that management aspects of organisational culture are also barriers. Staff fear reporting fatigue in case it results in a medical assessment (Filtness and Naweed, 2017). Instead they may phone in sick but not declare fatigue (Filtness and Naweed, 2017; Young and Steel, 2017). In addition, management messaging around fatigue can be inconsistent. Rosters can be designed to reduce the risk of fatigue, yet managers may then agree to shift swapping or overtime that leave inadequate time to recover. Staff raised concerns that managers do not monitor actual shifts worked (Filtness and Naweed, 2017). Secondly, with this research successful application of MWL measures would require organisational acceptance.

In future, fatigue management in rail could benefit from joint responsibility between employer and employee, monitoring fatigue reports, and monitoring psychophysiological state (Young and Steel 2017) once measures are sufficiently mature.

2.4.4 Skills rules knowledge framework

Rasmussen's skill-rule-knowledge framework (Rasmussen, 1983, Rasmussen and Jensen, 1974) is very relevant to signallers because it relates to those in control positions (e.g. processing plants), and performance is linked to varying levels of familiarity with the situation or task. Rather than claim that an individual is an 'expert' in all situations, it focuses on the type of situation or task individuals face.

- At the Skill-based level performance is based on pre-existing learnt analogue patterns of behaviour. Behaviour is unconscious. This level works well with routine, familiar, non-problematic activities.
- At the Rule-based level performance is based on a stored set of rules, where a situation is recognised as a fitting type, and this in turn defines the action required. Behaviour is conscious after a situation is identified as unfamiliar or non-routine.

- At the Knowledge-based level performance is based on conscious analytical thought, applied to novel situations, where previous learnt responses are unsuitable.

These levels exist within each person. Experience built over time increases the range of situations that can be recognised at the rule-based level, and experience to drawn on for experts at the knowledge-based level to analyse a novel situation and determine a suitable response.

The Generic Error Modelling System (GEMS) presents an integrated picture of what types of human error occur at each Skill, Rule and Knowledge-based levels (Reason, 1990). At the Skill-Based level slips and lapses occur during routine familiar activities and tasks. Skill-Based errors precede the problem. Examples include to inattention such as omission following an interruption or mistimed checks. Rule-Based mistakes occur when conditions deviate from planned and inappropriate rules are applied. These ‘bad rules’ could be either an encoding deficiency of the problem leading to misdiagnosis, or an action deficiency of applying the wrong rule. Both Skill-Based and Rule-Based errors are ‘hallmarks of expertise’ (Reason, 1990 p.59) and skilled performance, which are more abundant than the final type of error. Knowledge-Based mistakes occur when an individual realises the situation they face is outside their repertoire of Rule-Based solutions. Examples in include failure to notice the absence of relevant features or the tendency to focus on the wrong features. Either lead to Knowledge-Based mistakes being made. Both Rule-Based and Knowledge-Based mistakes occur after a problem occurs.

2.4.5 Expertise

Expertise is relevant to this research both in terms of seeking a measure of experts’ MWL and elicits knowledge from experts as input to the two interview studies. Definitions of expertise are numerous (Farrington-Darby and Wilson, 2006) and vary across disciplines of research including experimental psychology, computer science and knowledge acquisition

(Hoffman, 2014). Whilst types of expertise can include physical, cognitive, or social (distributed cognition) (Rasmussen et al., 1991), in this research the focus is on cognitive expertise. Early research in expertise studied chess players' ability to perceive patterns (de Groot, 1965) and encode positions into larger perceptual chunks (Chase and Simon, 1973; Lenat and Feigenbaum, 1988). Since then, aspects of expertise have expanded to extensive literature in psychology on strategies, judgement, decision making and associated phenomena of cognition (Hoffman, 1998), and the study of experts in cognitive science research (Shanteau, 1992). In this research the interest in expertise is focused less on the mechanisms of *how* experts make decisions, and more acknowledging that expertise contributes to managing MWL, successful completion of tasks and may, therefore, influence individuals' physiological data.

The research takes the cognitive science view that experts are more skilled and competent than novices (Anderson, 2000). Attributes from across disciplines that contribute to expertise include extensive knowledge, superior cognitive mechanisms such as memory organisation (Glaser, 1987), memory capacity, perception of meaningful patterns, identifying exceptions, and faster problem solving (Chi et al., 1988). Experts demonstrate a rich repertoire of strategies, are greater at inferring the meaning and implications behind information (Cellier et al., 1997) including dealing effectively with rare or tough cases (Hoffman et al., 1995). Experts have highly developed attentional abilities, can adjust decisions continuously, and have self confidence in their decision making (Shanteau, 1992) with economy of effort (Hoffman et al., 1995). Attributes relevant to railway staff include: inferring from information the implications for railway operations; faster problem solving so staff continue to meet the timetable where ever possible; self-confidence means they can make key operational or safety decisions independently; economy of effort; and a rich repertoire of strategies including tough cases; and the ability to adjust decisions continuously, enables staff to dealing with disruption or incidents effectively.

Experience built over time is an important contributor to expertise (Anderson, 2000; Bullough and Baughman, 1995; Ericsson and Smith, 1991). In this way expertise is not static (Bullough and Baughman, 1995), instead expanding and changing over time. The development of expertise can be broken down into five stages: novice, advanced beginner, competent, proficient, and expert (Dreyfus and Dreyfus, 1986). The concept of progression in expertise is not new, with craft guilds in the Middle Ages having a "Guild" terminology for development stages of expertise including novice, apprentice, journeyman, expert, and master (Hoffman et al., 1995). A distinguishing point when an apprentice becomes a journeyman is when the individual is deemed competent to perform a day's work unsupervised, working under orders. In rail this is equivalent to when a signaller or driver is passed out to work without a trainer present. Then experts are distinguished journeyman with extensive experience, who can deal effectively tough cases with economy of effort. A master is qualified to teach, and their judgements set the regulations or ideals (Hoffman et al., 1995). This research includes input from a range of experts from 'journeyman' experienced staff to those 'masters' who train them. The research has implications that extend to from trainees (apprentice) to master and the decisions made by managers and policy makers on the workload of staff.

Expertise is pertinent to this research in terms of methodological approach, as the research includes eliciting knowledge from domain experts. The robustness of such research relies on both identifying experts for participants and selecting appropriate methods for knowledge elicitation. Identifying experts includes consideration of their level of expertise such as years of experience, type of expertise including qualification, professional memberships, and whether they are practitioners with experience of daily problem solving or academic experts with more theoretical understanding (Shadbolt and Smart, 2015). The appropriate selection of which level and type of expert will depend on the knowledge being sought. In this research the focus is on practitioners, with some input from academics, but primarily qualified railway staff.

Various methods can be used with experts for knowledge elicitation. The appropriateness of these methods depends on the type of expert the type, and the type of knowledge such as Domain (declarative), Inference (concepts), Task (goals and procedures) and Strategic (broader system and controls) (Shadbolt and Smart, 2015). All four types can contain both explicit knowledge gained from what is taught and implicit, or tacit, knowledge from experience. Examples of methods that work with experts include observations, ranking exercises, interviews, and event recall. Observations suit identification of aspects of the task and procedures, and more implicit tacit knowledge learnt from experience (Milton N, 2003). Ranking or sorting exercises are used in some domains to explore hypotheses, whilst interviews work well to gain an overview of the domain, concepts, and reasoning (Hoffman et al., 1995). Semi-structured interviews add some structure to this method to make efficient use of the experts' time. One issue to note with experts in interviews is their differential access due to internalisation and reduced verbal access to their knowledge. As expertise increases, some aspects of expertise are not available to consciousness (Kim and Courtney, 1988; Salter, 1988). This means an individuals' knowledge increases, their ability to verbalise their knowledge decreases as conscious access to that knowledge decreases. Interviews may, therefore, yield more information from participants with intermediate levels of experience (Shadbolt and Smart, 2015). In an interview, or in addition to interviews, use of test cases can be an effective technique for knowledge elicitation (Grover, 1983). Event recall can also be an effective way to elicit knowledge as experts often have clear memories for tough or salient cases (Kolodner, 1991; Slade, 1991). All these methods are relevant to the research.

2.4.6 Decision Making

Klein's decision-making Models and frameworks in the literature explain human cognitive capabilities, such as decision making in complex conditions. Klein's Recognition-Primed Decision model (Klein 1993) indicates experts use their experience to form a repertoire of patterns enabling them to choose a suitable course of action rapidly. This would fit with psychology studies of visual recognition versus recall showing humans recognise images they have seen before.

In professional knowledge, Schön (1982) distinguishes between knowing-in-action, reflection-in-action, and reflection-on-action. Knowing-in-action includes 'Spontaneous behaviour of skilful practice' (p.51) and 'knowing more than we can say' (p.51). When performance matches expectation, humans tend not to think about it. This fits with Rasmussen's skills level, where experts complete their actions successfully without conscious awareness. Reflection-in-action 'is both a consequence and cause of surprise' (p.328), occurring within the 'action-present' whilst action can still make a difference. Experts seek cues to a standard solution, moving from a stance of tentative exploration to one of commitment (p.102). This fits with the rules and knowledge levels, and expert being conscious of the decision. Schön notes the importance of timing. What is meant by this is, reflection-in-action is swifter so can be applied in a timely manner during a task. It can be dangerous if it tips into reflection-on-action as this could unintentionally delay a decision, as reflection-on-action tends to be a longer process.

These concepts here are of particular interest in this PhD. Experts may be swifter, yet still suitable, as they draw on a repertoire of patterns. Secondly, a limitation of existing subjective measures is they only detect the conscious aspects of workload, and not all cognitive performance is conscious. Finally, in interviews, similarly, individuals may be able to recall moments of surprise but less able to recall what they did for actions that were successful. This matches what methods are suitable for use with experts. The SRK levels of performance explain how consciousness of actions align with different levels of expertise.

2.5 Models of Mental Workload

Mental workload (MWL) is a multi-dimensional concept (Xie and Salvendy 2000, Wickens 2008, Sharples and Megaw 2015) with no single agreed definition. Debate continues as to what exactly it is (Pickup, Wilson, Sharples, Norris, Clarke and Young 2005) and whether it measures what people do or how they feel about it (Pickup et al., 2005a). It is important to recognise the variation that exists in the definition of MWL. This presents a challenge when describing precisely what does or does not indicate MWL in previous, and this current, research. In this research, MWL is the umbrella term referred to throughout the thesis that acknowledges MWL is best understood through its constituent factors. A model of MWL is presented in Figure 2-7.

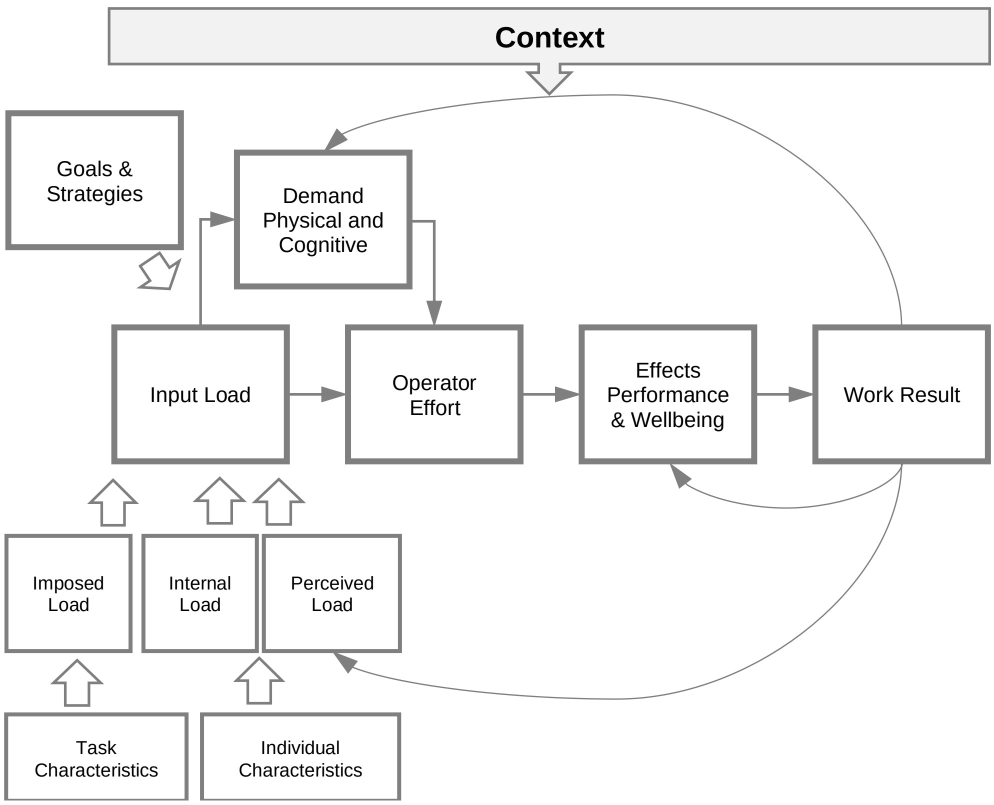


Figure 2-7 Mental workload Framework (Pickup and Wilson 2007)

Commonalities across this and other models and theories include a distinction between external *task demand* imposed upon a person, *individual* workload factors including effort and how they perceive the workload, and the *performance* outcome or work result. This research explores what specific factors of MWL could be inferred from physiological data, drawing on previous research findings and the novel contributions presented in this research.

External task factors are *stressors* such as, but not limited to, time pressure (Hendy 1997) and task complexity (Hart and Staveland, 1988). Internal factors include pre-existing individual characteristics such as experience, fatigue (Klein and Malzahn 1991), and other performance shaping factors (Kyriakidis et al., 2015). There are then information processing limitations that should be considered determining the mental capacity spent on the task (Kahneman 1973). The workload experienced by an individual can be measured in terms of *individual effort*, and any associated strain (Young et al., 2015). The NASA TLX workload measures, described in 2.6.1 , includes 'frustration' as a factor to include stress or annoyance. The inclusion acknowledges that emotional response can contribute to workload (Meshkati et al., 1995). In a complex work setting, Xie and Salvendy (2000) mapped out these external, internal and degrading factors, with their combination predicting performance (Xie and Salvendy, 2000), see Figure 2-8.

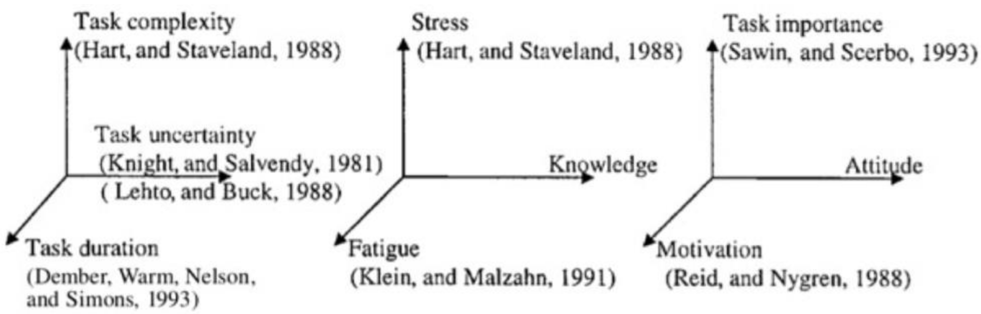


Figure 2-8 Factors contributing to mental workload

External, internal, and degrading in a complex setting (Xie and Salvendy 2000)

These models suggest various factors can individually, or in combination, impact workload and are relevant to understanding the workload of staff in live railway operations. Both task factors and individual factors will be considered in this research.

2.5.1 The dichotomy of overload and underload

Overload and underload are relevant to this research, as industry concerns. Regarding overload, as task demand increases, task related effort can initially be increased to sustain a level of acceptable performance. This is shown in region A3 the model in
Figure 2-9 (deWaard 1996). If workload continues to increase, overload occurs and performance decreases (region B in
Figure 2-9).

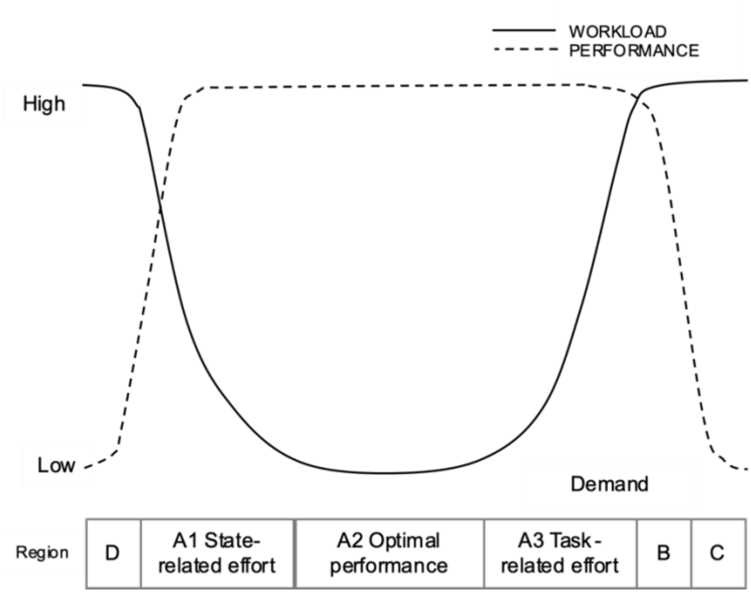


Figure 2-9 deWaard's workload and performance model (deWaard 1996)

When task demand increases, and effort can no longer sustain performance, research in air traffic control suggests a 'precipice of performance' is followed by a rapid, rather than graceful, degradation in performance (Edwards et al., 2016). This decline is as shown by the grey dotted line in Figure 2-10.

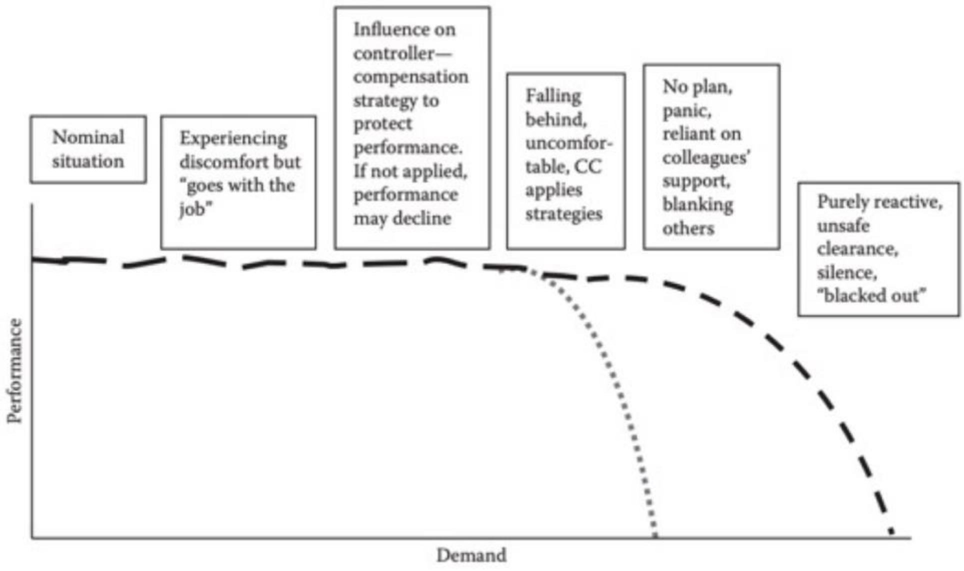


Figure 2-10 Degradation in performance with increase in demand (Edwards et al., 2016)

Regarding underload, deWaard (1996) also predicts that if workload drops to a very low level, then performance will drop once the operator cannot counteract their reduced state. On the other hand, Young and Stanton (2002) propose that performance decrements that occur with mental underload could be explained by reduced attentional resources. In a simulation car driving study they found, measuring secondary task performance, that total attentional resources shrunk to match the task demands. Whilst no studies were found that replicate these findings in rail, it seems likely that there are parallels relevant to a driver's or signaller's task.

In addition to understanding the range of potential contribution factors is the concept of how workload can accumulate over time (Xie and Salvendy 2000). Performance can be maintained during a temporary increase in task demand through increased effort. A sustained increase in task demand will result in overload and associated drop in performance (see Edwards' precipice of performance in Figure 2-10). This presents the possibility, in future, of determining if a pattern or chronology of factors lead to certain workload or performance outcomes. If a pattern in workload is discernible, and monitored over time, this could be a leading indicator of when workload is changing towards a negative impact on performance. This could apply to both overload and underload. In the rail industry this would be relevant to drivers and signallers. Underload risk could be assessed by considering attention. Overload risk could be assessed by considering duration of high task demand or other MWL factors.

2.6 Current measures of Mental Workload in the rail industry

The measurement of MWL has been the subject of research for over 50 years (Moray 2008). A wide range of measures have been established including primary and secondary task, subjective ratings, and physiological (Meshkati 1995). Originally MWL measures focused on discrete tasks (Moray 2008) in controlled conditions. More recent work considers dynamic tasks (Xie and Salvendy) and how individuals adjust their effort to sustain performance (Edwards et al., 2016, Young and Stanton 2002). Building on this work, and most relevant to rail, are measures that suit continuous dynamic tasks, detect a range of individual experience, levels of demand and effort, and identify how this experience changes over time, in live operations to measure Work-As-Done.

2.6.1 Criteria for workload measures

Selecting the most suitable workload measures requires consideration of the several criteria (Eggemeier et al., 1991, Sharples and Megaw 2015), as presented in Table 2-1. In this research minimising intrusiveness is a priority to ensure measures suit implementation in an operational environment and can be collected over longer periods. For this reason, secondary measures, that draw attention away from the main task, are not suitable for live operations due to their intrusion into the task and their risk of distraction. Sensitivity to temporal changes in MWL is also a priority, to capture the full range of MWL experienced in live operations. Staff acceptability is essential to apply these measures. Implementation considerations are relevant as they have practical implications for the researcher working in a live environment. Reliability can be drawn from the literature where evidence exists. Validity and diagnosticity are to be determined from the literature and the contributions of this research.

The criteria in Table 2-1 is narrower in their definition of MWL compared to the models of mental workload presented in Section 2.5 . Here selectivity/validity is deemed acceptable if only cognitive demands are detected as opposed to emotional stress. This is despite the models acknowledging that stress is a factor that contributes to MWL (Hart and Staveland, 1988). To determine the suitability of physiological data for the measurement of MWL. It is proposed here that that what factors physiological data are sensitive and diagnostic of should be explored first. The research remains open to what factors non-intrusive measures are sensitive to. Then an informed decision can be made as to whether physiological measures are beneficial to the measurement of MWL in live operations, and what defines MWL as a construct.

Table 2-1 Suitability of mental workload measures

Area of Expertise	Role
Sensitivity	Detect changes in task difficulty or demands
Reliability	Must reflect consistently the mental workload
Selectivity/Validity	Sensitive only to differences in cognitive demands, not other variables such as physical workload or emotional stress
Diagnosticity	Identifies changes in workload variation and the reason for those changes
Implementation	Includes aspects such as time, instruments, and software for the collection and analysis of data.
Intrusiveness	Should not interfere with the primary task performance
Subject acceptability	Subject’s perception of the validity and usefulness of the procedure.

This scoping review presents current workload measures used in rail, then the potential physiological measures that could fill gaps in the research.

2.6.2 Subjective measures of Mental Workload

Subjective measures capture individual experience, with high face validity, as data is derived from those completing a task. A task may result in different mental workload ratings for different individuals (Xie and Salvendy 2000, Pickup et al., 2005a, Matthews et al., 2014). Whilst this could be viewed as a weakness in a measure, Rouse *et al.* (1993) suggests this individual experience of load may be of greater relevance than imposed load. To give changes during a task these measures can require interruption of a task. Alternatively, an estimate can be given prior to a task or an average provided after a task. A Toolkit of measures was developed for rail in the early 2000s to suit use in the field for signallers (Lowe and Pickup, 2008), and drivers (RSSB 2005a, RSSB 2005b). Pickup noted in their Mental Workload Framework that workload comprises a sequence of stages. The measures developed addressed different aspects of this Mental Workload Framework.

Predictive subjective measures

The Operational Demand Evaluation Checklist (ODEC) is predictive tool, designed for signallers, that assesses potential task load and demand (Pickup and Wilson 2007). It counts the number of operational infrastructure elements such as junctions, stations, the volume of traffic in the timetable and complexity introduced by known incidents or failures. It can be used prior to other workload measures. It is extensively applied in the field by Network Rail and Human Factors consultancies (Delamare et al., 2016). Whilst it provides a first good “pass” on projects (Delamare et al), as technology changes, Network Rail are looking to update it (private communication 2018). As ODEC focuses on task demand factors, rather than individual experience, and Work-As-Imagined, it is out of scope in this research.

Subjective measures applied during the task

The Integrated Workload Scale (IWS), developed for signallers, rates subjective workload on a 9-point scale (Pickup, Wilson, Norris, Mitchell and Morrisroe 2007). The principles of IWS came from the Instantaneous Self-Assessment (ISA) tool (Tattersall and Foord 1996), a single 5-point scale designed to report a level of workload at regular points during a task. Originally designed by NATS (National Air Traffic Services). IWS, like ISA, was specifically developed for use in an applied setting. It collects real-time changes in MWL through a task, and the range of workload including peaks and troughs (Pickup et al., 2005b). It is suitable for use with both signallers and drivers in simulators to provide patterns of workload during dynamically changing work conditions. It does not provide information on the sources of workload. An example application was in a simulator, signallers gave frequent verbal rating of their workload which were recorded by researchers (Balfe et al., 2015, Pickup et al., 2005b). To date, if applied in live operations, the scale is verbally reported to an observer (Network Rail Ergonomics Team communication 2021). Despite verbal reporting, the risk of interfering with the task is sufficient that use is currently limited to simulator.

Retrospective subjective measures

These measures avoid interrupting the task, so suit application in simulator and live settings. NASA Task Load Index (TLX), (Hart and Staveland 1988) is the most widely used subjective measure presented here. NASA TLX has six subscales: Mental Demands, Physical Demands; Temporal Demands; Own Performance; Effort; and Frustration. It considers both demands from the task, and an individuals' feelings about it, identifying how much each factor contributes to workload. It does not provide temporal sensitivity. It has been used in rail, with adaptation, in research with train drivers (Large et al., 2014; Larue et al., 2016) and signallers (Thomas-Friedrich, 2017).

The Defence Research Agency Workload Scale (DRAWS) is another retrospective tool for train driver mental workload (RSSB 2005a, RSSB 2005b), based on a Defence Research Agency (DERA) workload tool (Jordan et al., 1995) to identify the nature of the workload being experienced, with four descriptors deemed more applicable to driver workload: receiving information, mental operations, making responses and time pressure.

The Adaptive Subjective Workload Scale (ASWAT) is designed for signallers to complete after a shift and was designed to compare two different times. It rates three factors that contribute to workload: time (how much spare time they have), mental effort (amount of mental effort or concentration) and pressure (level of problems, frustration, or anxiety). ASWAT was adapted from the Subjective Workload Assessment Technique (SWAT) (Reid and Nygren, 1988). Two main adaptations were, firstly, the term 'pressure' replaced 'stress', as signallers viewed stress as a weakness, whilst pressure was more frequently associated with workload than psychological stress. Secondly the weighting phase was removed to make the measure swifter to use in the field (Pickup 2006). The ASWAT provides an indicator of 'typical' workload at a location.

Overall subjective measures are sensitive and valid measures of MWL. Their sensitivity to temporal changes in a live environment however, come with a risk of task interruption. They can only provide part of the picture.

2.6.3 Observational

Observations, including time occupancy calculations (are used in live rail operations, capturing Work-As-Done, to assess proportion of time staff are spending on types of activity (Sharples et al., 2011, Balfe et al., 2008, Delamare 2016, Thorne and Rawlinson 2021). In such studies of signallers, the research records the predominant activity in the preceding 5 seconds, applying one of five categories: interaction (inputs e.g. setting a route);

planning (referring to timetable and live running information); monitoring; communication (telephone, radio or face to face); and quite times (e.g. non-work talking with colleague).

Observations benefits from minimal or no interference with the task but requires an observer and a Subject Matter Expert (SME) to help articulate the motivations behind the observed behaviours (Delamare 2016). A limitation is it may fail to detect MWL associated with monitoring, planning and decision making as individuals as theses may not present with physical observable markers. The data gathering requires a high level of attention from the observer(s), so it can only be applied for short periods (less than 1 hour). This in turn limits the range of workload that can be detected, as it is not possible to predict clearly which hour MWL will be at its highest.

If new measures could collect data over longer periods, they could show a wider range of workload throughout and across shifts. During COVID-19 all forms of this type of in-person assessment stopped, increasing the interest and relevance in measures that could be used without an observer present. In future, as tasks increase in monitoring, there will be a need for measures that determine workload from fewer observable behaviours.

2.6.4 Modelling

Computer modelling, and high-fidelity simulators, can predict an estimation of task demand. A measure for signaller workload is the D-MOD modelling tool (Delamare et al., 2016) provide both static and dynamic results of predicted workload for a signaller workstation. It incorporates a semi-automatic calculation of ODEC. The second signaller measure, currently being developed with Network Rail is the Workload Assessment Calculation Tool (WASCAL) (Zeilstra 2021). Modelling has also been applied to train diving (Hamilton and Clarke, 2005), including predictions of SPAD risk. A SPAD is when a train passes a red stop signal and is a potential precursor to a railway accident. In the United States it is known as a stop signal overrun. A strength of modelling is that multiple scenarios can be assessed, including normal running and perturbation. A limitation is they do not incorporate the varying demand from telephone calls. Modelling, as with any predictive tool, measures Work-As-Imagined, and inputs loads, rather than individual experience of workload. Ideally any measures that suit live operations could inform more accurate modelling.

2.6.5 Potential of physiological measures

Physiological, or psychophysiological, measures infer MWL from bodily activity such as heart rate. They show interesting potential to explore individuals' experience of MWL, including momentary peaks and underlying physiological state, with minimal task intrusion. These measures have been developed in parallel to MWL measures, and been applied in other industries, but with limited application in rail. When the research commenced, only two studies were identified from *live* railway operations: one involving train drivers (Song et al., 2014); and one involving signallers (Broekhoven 2016). During the research, psychophysiological metrics were noted by Rail accident investigators as a valid measure of drivers' physical state in future, specifically fatigue (Young and Steel 2017). A research gap was identified to explore how physiological measures could contribute to MWL assessment in rail. In doing so this research explores the physiology task interface, and the HCI of devices.

2.7 Conclusions

This chapter presented an overview of the rail industry context and a scoping review of the theories that underpin human cognitive performance, and the measurement of MWL. It provides a summary of the research and background on two research questions: How can temporal physiological data from wearable measures contribute to MWL assessment in rail industry live operations?; and what are the theoretical implications of individual physiological data to changes in MWL in a workplace setting? It identifies current MWL measures used in rail and theories of MWL that are relevant to MWL in a workplace setting. The scoping review scope included both drivers and signallers (controllers or dispatcher roles in other parts of world). Proportionally less research was found that focused on signallers, which presented a research gap. The research therefore went on to focus on signallers. The tasks described here reflect this focus, with detail provided on signalling. Whilst the focus is on signalling centres, the implications of this research are applicable to other control setting where staff have safety oversight and responsibility for operations.

The industry challenges identified were underload (attention) and overload (task demand and time pressure) as detrimental to performance. In addition, various automation technologies introduced to benefit staff workload may increase the risk of underload and increase workload during periods of disruption. This identified an opportunity for this research to explore the range of MWL, including between underload and overload. Taking this Safety II 'what goes right' perspective means measuring MWL that underlies sustainable successful operational performance. If new MWL measures could detect patterns, or cumulative MWL over time, these could provide both protective factors and leading indicators of deterioration in performance as MWL moves into underload or overload. Ideally applicable would be in the live operational environment to detect the full range of MWL experienced by staff. Such data could provide valuable feedback across levels of the socio-technical system. The challenge for a measure for the live environment is to be sensitive to MWL whilst minimising interference with the task.

Models help explain the different parts to MWL applicable to workplace, namely task demand, individual workload factors and performance outcome. In addition, research confirms various factors can influence MWL including external factor such as task demand and time pressure, and internal factors such as effort, alertness/fatigue, expertise, and self-confidence. Current MWL assessment in rail includes subjective measures, observation, and modelling. Each have different strengths and different limitations in terms of what they are sensitive to and how applicable they can be to live rail operations. Observation detects changes over time, with minimal task interference, but is limited to 1 hour and may not detect monitoring or planning as they may lack observable markers. Modelling and simulators can assess task demand but are limited in their realistic range of MWL. Subjective measures provide differences in individual experience, and changes over time, but are more intrusive to the task and may not identify the source of workload. The topic of whether a MWL should detect external task related changes, or internal individual changes is an ongoing debate.

Physiological wearable measures provide an opportunity to detect changes over time, with minimal task interference compared to asking staff to complete scales whilst working. They could inform management decisions regarding the management of MWL and the impact of change. They show potential to depict individuals' experience of MWL, including momentary peaks and underlying physiological state. In the rail industry, such measures would be applicable now and in the future in rail, with relevance to both current level of automation in the rail industry and ERTMS in future. Regarding what physiological data is diagnostic of MWL depends in part on the construct of MWL being ill defined as to whether it is only cognitive demands (as in table of suitability of measures) include contributing factors such as stress (as in Xie and Salvendy's research). This research is an opportunity to explore what physiological data are sensitive to in this setting (whether task demand or other factors), to then determine its suitability as a MWL measure. This in turn can inform the ongoing clarification on the construct MWL.

Chapter 3: Study 1 – Industry interviews on human performance in rail

"Nobody goes out there to have an accident, nobody goes out there saying "I'm going to mess up the system today". They're all just trying to get the job done and go home at the end of the day" (Participant 11)

3.1 Chapter overview

This chapter presents the results from Study 1, an interview study with industry stakeholders. The study explored rail challenges that relate to human performance or new technology, individual attributes of performance, and how performance is assessed in rail and other transport industries. The study sought to identity who impacts rail operations, and industry priorities of who to focus the research on. This study informed the subsequent research focus of scope on railway signallers in live operations in later stages of the research.

3.2 Introduction

Demand for rail travel has been increasing, with passenger journeys in 2017-2018 up 28% compared to 2007-2008 (Office of Rail and Road (ORR) 2018). To meet this increase in demand, the rail industry is increasing the capacity of the rail network by allowing more trains to use the tracks at any one time. To achieve this, automated and assistive technologies are being introduced to support some staff performing safety critical tasks. These staff have responsibility for the safety of themselves, colleagues, passengers, and the public. The industry seeks new measures to assess the impact of these new technologies on human performance in rail. Of particular interest are those technologies that assist or automate aspects of task. Their impact is that tasks are increasingly cognitive rather than physical, as more automation is introduced. This making the effort required more difficult to assess through physical measures or observation (Sharples et al., 2011). There is a risk is that managers of staff that see staff doing less assume that the task is easier when in fact, an irony of automation, the difficult parts of the operator's task remain or are more difficult (Bainbridge, 1983). The challenge is how to measure human performance and assess the impact of these new technologies.

Automotive and Air Traffic Control (ATC) industries are included here to identify whether human performance monitoring and assessment of road drivers or air traffic controllers could be applicable to train drivers and signallers. These industries were chosen for their similarly increasing automation and more cognitive, less physically, demanding tasks in control. Advancements in technology provide an opportunity to assess these changing roles in new ways. Current self-assessment workload measures require interruption of a task or application only the after completion of a task (Sharples and Megaw, 2015). Physiological measures offer the potential for continuous data without interrupting the task. Physiological measures detect aspects of physical activity, such as Heart Rate Variability (HRA), or Galvanic Skin Response (GSR), that can be used to infer levels of cognitive activity. This study considers whether such technologies could be applied in future to assess human performance in rail. An important consideration is how the data from these technologies will be used and whether they can fit with wider competence and performance assessment processes.

The purpose of the study is to investigate rail industry challenges relating to human performance, and how data on human performance is currently assessed across three transport industries. The study applied a pragmatic approach by seeking stakeholders' perspectives to identify current challenges and assessments and inform the research focus. As perspectives of stakeholders are likely to shape their acceptance, understanding these industry perspectives can guide the subsequent research towards a topic pertinent to rail that the industry will both engage with and benefit from. The first part of this aimed to identify whose performance to measure, including any specific new technology. The second aimed to explore the types of current data collection of human performance, and what types of measure could be applicable to rail in future. The findings from this study, in combination with the scoping review, informed a focus that is a research gap and is perceived as pertinent the rail industry.

3.3 Method

3.3.1 Study design

The study applied a pragmatic approach (Robson and McCartan, 2015) by seeking stakeholders' perspectives to guide the subsequent research focus towards a topic pertinent to the rail industry. The value of the findings was in identifying what works for industry, recognising that the reality of the rail industry's operational setting is multiple, complex, constructed and stratified (Reichardt and Rallis, 1994). This initial study took a broad and high-level scope to gain a wide range of experience and opinions from stakeholders across transport industries. Findings were used to contextualise the findings and develop and refine the research questions (Thompson, 2017).

This study used semi-structured interviews to explore how human performance is currently assessed in rail, and how it could be measured in future. The use of semi-structured interviews allowed the stakeholders to share their experiences and opinions in their own words (Coveney, 2014). The method included elements of both an inductive and deductive approach. The semi-structured interview prompts supported a top-down deductive approach (Robson and McCartan, 2015) with stakeholder providing answers and examples on existing topics. The bottom-up inductive approach was applied as the interviews progressed and additional topics emerged beyond the original prompts (Braun and Clarke 2012). The coding and analysis stages used a combination of both inductive and deductive approach to establish themes. Themes reflected both the original deductive prompts and emerged during the interview process and coding. The themes that emerge from the interview data go beyond what can be observed (Glaus et al., 1996). The final themes and sub-themes reflect those relevant to the rail industry to inform the focus of subsequent stages of the research.

3.3.2 Participants

Semi-structured interviews were conducted with 14 participants, representing stakeholders from across the rail industry and transport industries. Stakeholders were those in roles most likely to make use of human performance data, with a vested interest in any new measures that are developed. They are the ones who have the potential to influence the future uptake of measures. The range of stakeholders included: those with current or previous front line operational experience; managers of staff or operations who are the users of the information gathered on human performance to inform operational decisions; and experts, including Human Factors experts, who are the gathers of data on human performance and developers of new measures to inform industry. A full list of stakeholders is presented in Table 3-1.

Recruitment was through the two industry organisations, Railway Safety and Standards Board (RSSB) and Network Rail, both industry sponsors of the research. A snowball sample approach was then taken, with each participant being asked to suggest other participants to ensure a range of perspectives from across the industry. The stakeholders from other industries were approached through professional contacts of researchers at University of Nottingham. Participation was voluntary and no financial incentive was provided.

To provide anonymity, all interview responses were labelled with a participant number. Participants provided a description of their role. The range of roles and expertise of these stakeholders are presented in Table 3-1.

Table 3-1 Participant roles and expertise (Train Operating Company (TOC), Freight Operating Company (FOC)

P. No.	Area of Expertise	Role
P1	Train Operations (TOC)	Head of Operations, Ex Driver
P2	Railway	Human Factors, Specialist in Human Performance
P3	Railway	Risk Expert
P4	Signalling	Signaller
P5	Signalling	Human Factors Expert
P6	Signalling	Human Factors Expert, Ex Signaller
P7	Train Operations (TOC)	Head of Operations
P8	Railway	Human Factors Specialist
P9	Train Operations (TOC)	Head of Drivers, Ex Driver
P10	Automotive Industry	Human Factors Senior Academic
P11	Railway	Accident Investigator
P12	Air Traffic Control	Head of Human Factors
P13	Train Operations (FOC)	Operations Standards Manager, Professional Head of Operations, Ex Driver
P14	Rail Simulation	Systems Modelling and Simulation Expert

3.3.3 Procedure

The interviews were semi-structured, consisting of open-ended questions in two categories: current challenges related to human performance; and current data collected on human performance. During the interviews broad initial questions were asked including: what current challenges in rail rely on the performance of humans?; what new technologies are involved?; what data currently exist in rail that capture human performance?; how are data collected?; and who uses the data? Follow on questions were asked to encourage stakeholders to explain their answers and give examples. In the final part of the interview, participants were provided an opportunity to recommend other people to be interviewed to provide another perspective. The study received ethical approval from University of Nottingham, as shown in Appendix A. The protocol used for open questions is also presented in Appendix A.

The interviews were conducted between March and October 2018. Eleven interviews were conducted face to face, and four over the telephone. All interviews were audio recorded, with notes taken during the interview to guide questions and capture themes. The interviews were planned to last 45 minutes, with the actual interviews ranged between 29 – 87 minutes and averaged 54 minutes.

In addition to the interviews, a TOC meeting of operational managers was attended which covered driver training, recent incidents, and current operational issues. This provided operational the context in which to situate the subsequent themes that emerged through the coding of the interviews.

3.3.4 Data analysis

Interview data was analysed in a series of steps. Firstly, the interviews were transcribed verbatim. Secondly emerging thematic analysis (Strauss and Corbin 1990) was used to identify codes and broad themes. An initial coding template was built by coding on paper using coloured pens and applied to five interviews. During this stage, fourteen initial codes emerged. The third stage involved transferring the coding from the initial paper copies into NVivo where another iteration of coding was conducted (Saldaña 2016). This iteration grouped the initial codes into themes and sub themes. The resulting themes were then applied to the remaining 9 interviews. The final themes are presented in Figure 3-1.

To further explore the theme of ‘individual attributes’, a comparison was completed between opinions offered by stakeholders of what indicates ‘good human performance’, with the evidence base of two existing frameworks: RSSB’s Non-Technical Skills (NTS) (RSSB 2012) and Risk-Based Training Needs Analysis (RBTNA) tool (RSSB 2018). These highlighted which opinions match existing frameworks in industry, and which were additional.

3.4 Results

The thematic analysis of the interviews identified four key themes relating to human performance in rail: People who impact rail operations; Time when assessed; Individual attributes; Future types of monitoring.
The coding tree is presented in Figure 3-1. Results are then presented for each theme.

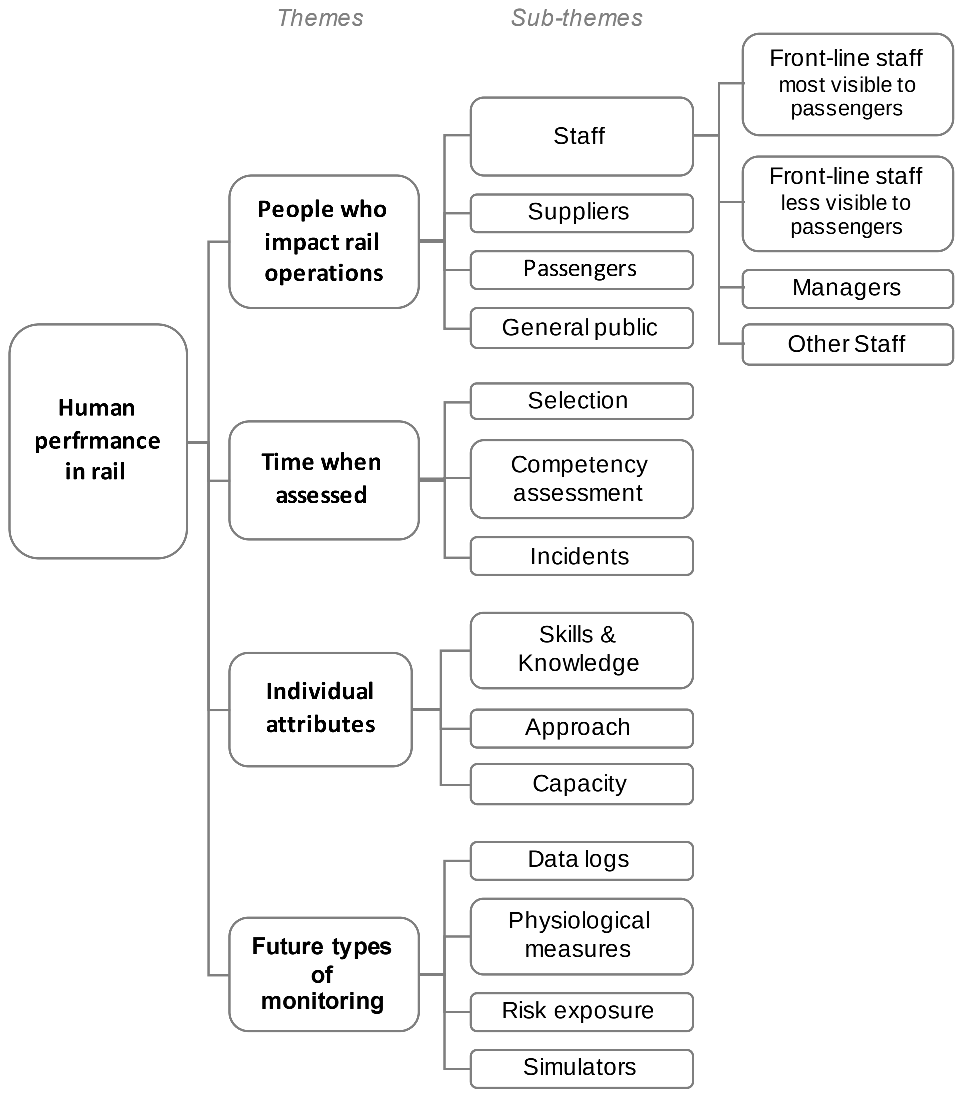


Figure 3-1 Coding tree of key themes from industry interviews

3.4.1 People who impact rail operations

A theme that emerged during interviews was that, rather than only staff, the performance of a broad range of people can impact rail operations. One specialist, when asked where in rail humans impact rail operations, answered "everywhere" (P11). The following people were mentioned in interviews, with categorisation added here based on their visibility to passengers and the industry definition of Safety Critical Tasks (ORR 2017):

- Staff
    - Front-line staff performing Safety Critical Tasks most visible to passengers: Driver; Guard; Train Dispatcher, Track workers.
    - Staff performing Safety Critical Tasks less visible to passengers: Signaller; Level Crossing Operator (including CCTV); Electrical Controller; Mobile Operations Manager (MOM); Maintenance Fitter; Shunter; Axel Inspector.
    - Managers: Managers of front-line staff; Standards Manager; Train Services Manager; Operations Manager; Head of Engineering; Head of Safety; Station Controller (who allocates platforms); Shift Signaller Manager; Track Section Supervisor; Train Running Controller; Route Control Manager; Route Control Incident Manager.
    - Other staff: Timetable Planners; Control Centre Technicians (receive fault alerts from remote condition monitoring).
- Suppliers to operations: Signal Designer; Traffic Management System Designer; manufacturers; funders; government; Rolling Stock Leasing Companies (ROSCOs).
- Passengers: at the platform train interface
- General Public: road users at level crossing users; trespassers.

This wide range of people shows how complex a socio-technical system rail is. What became apparent during interviews is the performance of any one human in the system is interlinked with the performance of many others. In terms of prioritisation of research, stakeholders indicated drivers and signallers would specifically suit being the focus of further research. This was because drivers and signallers are two of the biggest groups of staff, have high levels of assessment, and complete Safety Critical Tasks. The difference noted between the two staff roles is the extent to which it is visible to passengers and the public, with drivers being a more 'visible' staff role. This is interesting as the scoping review of literature found that proportionally less research has been conducted on signallers. The less visible staff play an important role in safety and present a research gap for this research.

3.4.2 Timing of competency assessments

Comments revealed concerns about driver and signaller competence at different career stages. Assessment predominantly occurs at selection during recruitment, at routine competency review; and following an incident. Each will be presented in turn.

Selection

Selection during recruitment includes an interview, as well as medical and psychometric tests (drivers only). One TOC (P9 and TOC meeting) felt that the standard of candidates had reduced in recent years. More applicants now pass their assessment, but then fail basic training. The TOC suggested that this increase had been since a change to the industry's standard Initial Driver Assessment in 2013. The TOC had four drivers fail training on one course of 8 candidates (P9, TOC meeting):

> *"We've had more failures through training in the past few years than we've ever had and at least 25% of our incidents now are down to post qualified drivers"* [P9, Head of Drivers, Ex Driver, TOC]

There is also anecdotal evidence of a change in applicant demographic:

> *"Youngest freight drivers are late twenties, mid-thirties, early forties. Probably as a second career sometimes (having) been in the army, or police, or teachers. We've had a bank manager who's joined because it's good money… So I think the demographics are going to change over the years."* [P13, Head of Operations, Ex Driver, Rail]

> *"We lost people who had been in the industry for 20-30 years and spent the weekend looking at trains at the end of platforms. Then (new) junior people are doing it as a computer game in a sense, who have to learn the real world. It takes a long time for them to get to speed."* [P5, Human Factors Expert, Rail]

One specialist suggested the pay offer could be encouraging this change in applicants (P7) and implied that this did not improve the suitability of individuals for the role. These findings may indicate that what motivates individuals to apply to rail is changing, resulting in people with a wider range of previous experience and interests working in rail. Some individuals may have more intrinsic reasons for applying such as an interest in rail, others may be motivated by more extrinsic factors such as pay and working conditions.

The high failure rate of basic training suggests firstly a concern over the current accuracy of assessment techniques, and secondly that some individuals are more suited to the roles than others. Opinions expressed by stakeholders suggest that they believe that those with an enthusiasm for rail in their personal life are more suitable. Conversely younger 'gamers', or those from outside rail, are less suitable. Stakeholders also described experienced staff leaving the industry as a loss, implying it is a negative. Potentially the introduction of new technologies can lead to experienced staff choosing to leave rather than learn a new way of working. The same technology attracts new recruits who already work in a different way. There appeared, overall, to be a concern around how to manage the change, and a lack of clarity on what should be selected for.

Competency assessment

Once qualified, staff competency assessments occur on a rolling programme over 1 - 2 years (P7, P9). Signallers are observed at their workstation and drivers during a cab ride. Samples of voice recording are checked for correct use of safety-critical communications (P8). Driver monitoring includes samples of data from the On Train Data Recorder (OTDR) to check performance such as speeding (P7) and provide feedback on good performance (P9) e.g. correct sequence of actions (P7). Currently, signaller workstation control logs are not routinely reviewed (P6).

Two specialists (P3, P13) questioned whether staff perform differently when observed. To address this a concern a FOC has moved away from relying on observations:

> *"The primary method of assessing a driver is a download (from) data recorders because they (a driver) will drive nearly perfectly if someone's sat next to them … your data recorder assessment is probably warts and all. The detail you can get is quite astonishing."* [P13, Head of Operations, Ex Driver, Rail]

A FOC's safety team receive live OTDR alerts e.g. if a driver leaves a work site without turning the Train Protection Warning System (TPWS) back on before driving over 40 mph. Patterns in OTDR data can also show erratic actions such as braking and use of the Drivers' Vigilance Device[6] (DVD) (P13). This manager has seen in the data how drivers who are fatigued acknowledge the Automatic Warning Systems[7] (AWS) alarm in different ways. Drivers have 3 seconds to respond to the AWS. The manager gave the comparison between a driver who is wide awake will respond to reset it in one second, and a fatigued driver who will either respond to it at the very end of that time or may press it too early as pre-empting it "almost on auto-pilot" (P13). During a review of data from an incident, the manager reported noticing, over the course of the journey prior to the incident they could see the driver's reaction times increasing. The manager believed this indicated the driver becoming more tired over time. The manager also found reviewing snap shots of the data with the driver a useful part of their competency assessment, to

[6] The Driver's Vigilance Device (DVD) sounds an alarm after a period of driver inactivity. The driver must relieve pressure on the Driver Safety Device (DSD) pedal, and reapply it, in a short amount of time. If they do not the emergency brakes apply to the train.

[7] AWS gives drivers an audible warning and visual reminder they are approaching a distant signal at caution. If the driver does not acknowledge the warning the brakes are applied.

complement their observations in the cab. Whilst the detailed OTDR data is not available in real-time, if there is an event that warrants investigation, the data can be downloaded remotely, and a time-period reviewed.

Simulators are used for training, such as drivers' route learning (P9) and to practice degraded working (P13), however individual performance is not assessed during these sessions (P7).

Incidents

Staff are interviewed following an incident e.g., a Stop Short Door Release (when only part of a train is at a platform and passengers risk falling onto the track). Various sources of data can support further investigation. Investigators can check staff received everything they needed (P8) e.g., training. On Train Data Recorder (OTDR) data can identify driver errors or confirm their account of events. The quality and ease of access to OTDR varies across the industry. Signallers' inputs are logged, but only at modern electrical control centre workstations (P6, P8). These logs do not identify the individual so linking data to an individual requires checking their roster.

Factors that influence performance are also considered such as fatigue and levels of experience. Participants reported that drivers rarely report feeling fatigued (TOC meeting), although it is known to be an industry issue (P2, P13) that contributes to incidents (P13, TOC meeting). In terms of experience, one TOC estimates that 25% of their incidents are by drivers in their first year (P9).

Participants felt that there is an element of chance as to whether an individual has an incident (P2) and noted that individual performance has the potential to limit the risk of an incident occurring.

3.4.3 Individual attributes

Responses in interviews focused on the measurement of human performance through competency assessment and incident investigation. Individual attributes as a theme emerged over the course of the interviews. What made someone good at a role seemed difficult to define (P4). Participants were therefore asked what made a good driver as an example role. Their responses are presented in the first two columns of Table 3-2. Non-Technical Skills (NTS) were noted as important by stakeholders (see list presented in Appendix C). Where interview responses map to the existing industry NTS framework (RSSB 2016) is shown in the final column of Table 3-2. The skills and knowledge theme was noted as matching requirements identified in the Risk Based Training Needs Analysis (RBTNA) tool (RSSB 2018) (P2). This tool provides the industry with guidance to determine staff training needs. It was developed for the role of train drivers but is intended to be applicable to other staff.

Whilst physical attributes, gender and age were not mentioned, the application of skills, personality and attitude were mentioned. Expanding on the capacity for good concentration mentioned in Table 3-2, a driver manager gave examples of good or poor concentration based on their experience of observing drivers in the cab. A driver they had observed who seemed good at maintaining concentration had concentrated from station to station, taking each stage of the journey at a time. In comparison, drivers they had observed who seemed to find it difficult to maintain concentration were those who were distracted more easily, or their mind wandered. They suggested that the lifestyle of some individuals outside of work may leave them less able to deal with long periods of work, such as due to insufficient sleep.

Table 3-2 Perceived beneficial attributes for drivers

Attributes mapped to Risk-Based Training Needs Analysis (RBTNA) tool and Non-Technical Skills (NTS)

	Interview Responses	RBTNA tool	NTS
Skills & Knowledge	Excellent Non-Technical Skills	Non-technical Skills	All NTSs
	Practical skills e.g., train inspection	Functional Skills	n/a
	Knowledge	Underpinning knowledge	Self-management: Maintain and develop skills and knowledge
Capacity	Good situation awareness	n/a	Situation awareness: Overall awareness
	Good concentration, can deal with long periods of work	n/a	Situation awareness: Maintain concentration
	Can take on lots of information	n/a	Situation awareness: Retain information
Approach	Reliable e.g. turns up	n/a	Self-management: Motivation
	Methodical	n/a	Conscientiousness: Systematic and thorough approach; Checking
	Prepares & plans well	n/a	Self-management: Prepared and organised
	Takes each stage at a time e.g., concentrates from station to station	n/a	Workload management: Prioritising
	Confidence	n/a	Self-management: Confidence

In addition to the findings in Table 3-2, there were some comments that implied that personality or attitudes impact drivers' performance. A stakeholder believed that introverts made better drivers than extroverts, as they perceive introverts as self-contained and better at working solo. In terms of attitude, an 'every day's a school day' was viewed as a positive, whilst a 'what's in it for me' or a complacent attitude were viewed negatively. These findings indicate stakeholders can believe the capacity and approach of staff varies between individuals and that attributes can impact staff performance.

Demographics were viewed as quite superficial (P2). It was noted that experienced staff can find the transition to newer technologies, such as tablets, difficult (P13), with some choosing to retire rather than transition. A level of uncertainty was detected in interviews as to whether staff with experience of older ways of working, or newer recruits who only learn the newer technology, are better able to achieve good performance. One participant described younger signaller applicants as ‘gamers’ who have a different mindset, with the benefit of adapting quickly and positively to digital technologies (P14).

The number of incidents an individual had was not deemed a reliable measure of performance (TOC meeting). It is a point of debate for accident investigators how much previous incidents are relevant to a current investigation (P11). TOC operational managers expressed confidence in some drivers who had had an incident and a lack of confidence in some drivers who had not had an incident.

In terms of the attitude of staff:

“Nobody goes out there to have an accident, nobody goes out there saying “I’m going to mess up the system today”. They’re all just trying to get the job done and go home at the end of the day” [P11, Accident Investigator, Rail]

It was noted that demands for precision are increasing (P1), and staff are close to the limit (P5):

“The requirement for a train driver now is that you need to be 100% accurate 100% of the time and there isn’t any let up to that. Twenty-five years ago, there wasn’t the pressures in the performance, capacity, and the safety requirements” [P1, Head of Operations, Ex Driver, TOC]

Interviews highlighted challenges beyond individual roles. Firstly, due to how long technology deployment takes, staff must constantly operate with a mixed age of technologies. Secondly, across the industry corporate knowledge is lost when staff leave the industry (P5, P14) leading to decisions being made that impact front line staff by individuals with limited understanding of the implications. Thirdly, driver and signaller roles have become more mundane (P13): driver routes are not varied (P7), trains have become easier to drive, with more comfortable cabs (P1) and automation protects signalling staff actions (P4). These all have the unforeseen risk of leading to inattention due to underload (P2), (TOC operational meeting). One participant questioned how individuals can maintain focus (P7).

The industry faces a combination of challenges. How can individual factors, such as capacity or approach, be objectively measured to manage MWL effectively? Objectivity could assist in removing biases in interpretation, reduce associated cultural tensions, and encourage collaboration. Challenges include experienced individuals choosing to retire which reduces the pool of experience in teams that takes years to rebuild. Also, it remains unclear how close to their MWL limit staff are working, including the risk of underload.

3.4.4 Future types of monitoring

In this theme, stakeholders provided their perspective on where and how different measures show potential for future use in rail to assess human performance.

Data logs

Stakeholders from across the transport industries provided examples of ways systems could provide objective data to support assessment of performance. In the automotive industry, lane keeping is now available in some cars to monitor driving performance in real time (referred to as lateral driving performance) (P10). In ATC, objective data from mouse clicks, and time spent on the radio, can indicate the complexity of the task (P12). In rail, OTDR data can indicate different personal driving styles (P8). Stakeholders proposed OTDR data could be used (P3) to record an 'optimal' drive performed by an experienced tutor to show to trainee drivers (P13).

Physiological Measures

Results from the interviews indicate that the application of physiological measures is at different stages of development in different industries. The opinion of stakeholders was that the automotive industry was the most advanced, with measures currently available in some cars on the market. ATC has trialled various physiological measures in live operational settings. In rail, operational staff were not aware of any physiological measures being used in live operations.

The automotive industry is developing real-time driver monitoring features (P10) including cameras to detect aspects of the face including eye position to monitor slouching (P10). Products such as fatigue monitoring are available at the high end of the market such as a dashboard coffee cup indicator to recommend a driver takes a break when they show signs of fatigue (P10, Car Sales 2018). Features that are deemed to be for safety are generally well received by customers, although attitudes to being monitored vary between cultures and between generations (P10).

In ATC, trials of physiological measures found that they could monitor individual consistency and spot change. Using Electroencephalogram (EEG) for example, it was possible to detect when an individual's readings were different from that individual's normal reading. The next day the individual became ill. This demonstrates how change in EEG detected a physiological change before the individual was aware of becoming ill (P12). Other biometric data collected in ATC includes visual scanning patterns (using eye tracking) and how individual Controllers deal with stress (using a chest strap to detect heart rate variability, and GSR from a device worn on the arm) (P12). In addition to physiological measures, in ATC subjective measures are routinely used including workload, confidence, and Situational Awareness to assure the system is safe (P12). The stakeholder from ATC postulated that in future, in the live environment, data from physiological devices could provide the supervisor with real time indicators and a decision support tool for when the task load of a Controller needed adjusting. In ATC, the task load can be adjusted by rerouting aircraft to different sectors.

Operational staff in rail did not mention physiological measures being used to monitor staff. They instead reported competency management as the current way human performance of staff is monitored. Human Factors Experts were more aware of the potential of physiological measures. One trial that was specifically mentioned was one that found train driver GSR correlating with difficult conditions such as risk of train slipping in low adhesion conditions (P8) (Crowley and Balfe 2018).

Risk exposure rates

Two tools were provided in interviews that calculate 'exposure to risk' rates, and therefore likelihood of adverse events occurring. Firstly, the Red Aspect Approach Tool (RAAT) (P2) is a tool to indicate how frequently drivers are exposed to the risk of a SPAD (P1). It does this by determining how many red aspect (stop) signals a driver approaches on their route. A red aspect will indicate to a driver that a track section is occupied. A SPAD is when a train passes a red aspect signal, and therefore is at risk of collision with whatever is occupying the track ahead. The second tool would use existing timetable information to calculate the exposure rate of drivers to incidents of Stop Short or Failure to Call at a station, based on the number of times a driver stops at a station (P1). These incidents can occur, for example, when a driver drives a route where frequently they drive a long train (multiple carriages) that stops at only major stations on a route. When they then drive the same route as a stopping service, they are at risk of Failure to Call as they are used to driving through some of the station stops. If the length of trains they drive, they may become used to stopping part way along a specific station platform with a shorter train then in error stop at the same point when driving a longer train. This would then not allow passengers to safely disembark or board the train on the carriage as it is not level with the platform.

Simulators

Simulators in rail are currently used for training (P8), but not widely for competency assessment. Simulators provide both trainee drivers and signallers with opportunities to learn how to complete their tasks before applying their training in a live operational environment. Currently only about 5% of the data collected from a simulator session are used (P14). There is a potential that, however, simulators could be used to study aspects of performance such as reaction time (P14).

3.5 Discussion

The industry challenge for human performance in rail that emerged from the findings related to risks resulting from overload and underload. Concerns were raised that staff are approaching an overload limit as demands increase for precision and capacity whilst maintaining safety. Underload was a risk noted for the driver and signaller due to their roles becoming more mundane, making it difficult to maintain focus. A second challenge was, rather than a specific technology, the introduction of any new technology that impacts staff as they must adapt their way of working and work with mixed ages of technology. Any novel technology brings with it a period where staff are inexperienced in using that technology. An impact is that staff choose to retire rather than transition. A further impact is new replacement staff, with 25% of train driving incidents involving drivers in their first year. These findings in combination suggest that there is a need in industry to assess and monitor the impact of introducing new technology. Rather than only assessing a specific technology, or only at specific points in time, measuring human performance over longer periods to detect changes over time. This could provide data on periods of time where human performance is successful and seek to detect periods of overload or underload. Comparing to data before a change could assess the impact of new technology. Equally such data could be used as a learning aid for new staff as their build experience. These study findings and viable future measures are summarised in Figure 3-2.

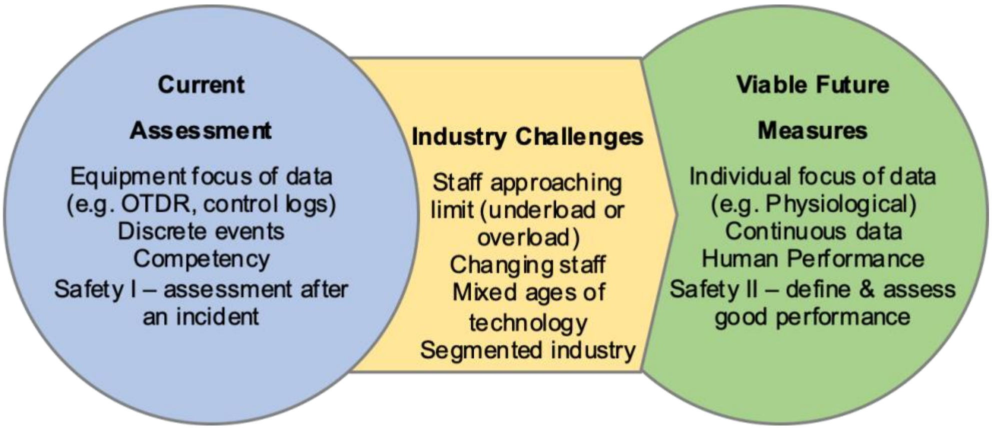


Figure 3-2 Challenges and future opportunities for human performance assessment in rail

The theme of people whose performance impacts rail expanded during the interviews to include not only staff but suppliers, passengers, and the public. This finding expanded the potential choice of whose performance in rail the research could focus on. It was determined that front-line staff completing Safety Critical Tasks, in particular drivers or signallers, would be a suitable research focus that would fit with industry priorities. The distinction between the roles is that the signallers are less ‘visible’. The scoping review discovered proportionally more research focused on drivers than signallers, indicating research into signallers as a suitable research gap. A focus on one role may be generalisable to other staff in rail and other transportation industries, particularly where similar risks of overload or underload apply or in other control roles.

The theme ‘individual attributes’ emerged from interviews as attributes stakeholders associated with individuals who deal well with the workload and achieve successful performance in their roles. Many match good rail Non-Technical Skills. In the light of industry concerns about reaching limits and levels of experience, two attributes stood out: confidence and concentration. Maintaining concentration is recognised in the industry as important, yet there is an industry concern about it being achievable. Confidence was mentioned in interviews as being measured in ATC and is also included in NTS. It is not currently measured in rail, so presents a potential research gap. In addition, some attitudes and personality types were implied by stakeholders to have an impact on human performance. Whilst some attitudes are implied in the self-management aspects of NTS, personality is not explicitly covered. It is of interest here, however, that stakeholders believe that there are aspects of successful performance that depend on individual characteristics and not solely on the external task demands. This suggests a benefit in acknowledging that individual factors are relevant if human performance is to be assessed.

The current assessment of human performance in rail is around competency and occurs at certain times. Data focuses on human equipment inputs (e.g. OTDR data), and less on individual performance. One challenge in rail is that as roles change, appropriate competency is changing without an associated change in assessment. If incidents occur, a more detailed data analysis is completed. This fits with the Safety I approach (Hollnagel, 2014), which tends to be reactive and can lack the awareness to predict incidents. As one participant said, rail has “no way of knowing how close to the limits of ability staff are” (P5). Adopting a Safety II approach in future would focus on identifying what factors contribute to good human performance and result in successful operational performance. These factors include the individual attributes of concentration, confidence, and identifying when staff are approaching an overload or underload limit.

In considering future monitoring, increased use of data logs (e.g. OTDR) and simulators are possibilities in rail. OTDR data shows potential for further use (Balfe et al., 2017), although Balfe notes there are currently no validated metrics that differentiate good from poor performance of either the drivers or infrastructure (Balfe et al., 2017). These would need to be developed. Additional use of simulators, and assessments of risk exposure also show potential. In terms of identifying a research gap that focuses on personal data, the use of physiological measures in rail is a topic for further investigation. ATC, rail, and automotive industries are exploring the potential of various physiological measures. These measures offer a potential future opportunity to collect and analyse continuous data in rail, without interrupting the task. This in turn will allow measurement of good performance and help to prevent incidents. It will be important to assess the suitability of these measures, practicalities of applying them, and ethics of collecting data that measure individual performance of staff. Further research should include consideration of the attitudes of staff to the use of such data, who would have access to it, and how such data could be used within wider competence and performance assessment processes.

The terminology used when discussing performance needs further thought. Whilst the phrase ‘human performance’ is not used in rail by operational staff, or seen as a distinct issue, participants did recognise that the performance of a wide range of people impacts rail, including passengers and the public. There were also examples of good work and research around human performance, but a lack of awareness of this across the whole industry. This fits with the finding that the industry is segmented, with priorities varying across operations, locations, and time scales.

3.6 Limitations

In aiming to identify a topic for the research scope to focus on, this study took a broad and shallow approach to the structured interviews. This constrained the opportunity to uncover and explore stakeholders' underlying reasons for their perspectives. The resulting coding structure tends to reflect categories and themes for further research of the 'who, what and how', rather than identify underlying more abstract concepts that reflect the perspectives of staff of the 'why'. Secondly, the study was the first by the researcher to apply a qualitative approach. Whilst this provided a learning opportunity for future studies, it constrained this study. Redressing these limitations in future could provide a richer picture of the reasons for stakeholder perspectives on challenges in rail, human performance data, and beneficial attributes in staff.

This study aimed to identify what measures could be used in future in rail. For this reason, the stakeholders primarily represented the rail industry, with only one stakeholder from the aviation and automotive industries. If a more even proportion of stakeholders were recruited in future, a comparison of perspectives across different transport industries could be conducted. This could be an interesting study, particularly to understand the relative difference in underlying organisational cultural maturity.

3.7 Conclusion

The study explored rail challenges that relate to human performance or new technology, individual attributes of performance, how performance is currently assessed, and could be assessed in future. The study gathered stakeholder perspectives and opinions, to guide the subsequent research towards a topic pertinent to the rail industry that the industry will both engage with and benefit from. The study identified underload and overload as current challenges in rail relating to human performance. Another challenge is that when technologies are introduced, staff are inexperienced in using the novel technology. An opportunity exists here to seek measure of human performance. This could measure successful performance as a baseline. Then the impact of new technology could be assessed against the baseline, or the risk of poor performance be predicted when MWL moves into overload or underload, or be used as a learning aid to support new staff whilst they build experience.

Regarding whose performance to measure, a wide range of people were identified as having an impact on rail operations. Of these, drivers and signallers were identified as the priorities for the industry. Of these two, signallers were chosen as they are the less visible role and present the larger research gap. Whilst research will focus on signallers, results from assessments of measures may be generalisable to other roles in rail and transportation industries, particularly control roles.

Currently in rail, performance is assessed intermittently, using mainly competency assessments. Future research opportunities exist to expand the use of data logs, risk exposure tools, and simulators. In addition to assessing the impact of task factors, there was acknowledgment in the interviews that individual attributes contribute to performance. Maintaining concentration, and confidence are two specific characteristics that present a research gap to consider when measuring human performance. This research will explore the use of physiological measures as they fit the personal data perspective of the research and show the potential to provide continuous data. Further work will be needed to assess their suitability for assessing a range of MWL with associated underload or overload, for unobtrusive use, and for the acceptance of staff to their use.

Chapter 4: Physiological measures for mental workload assessment

4.1 Chapter overview

This chapter provides a scoping review of physiological measures and the critique used to select the measures suitable for signallers in live operations. The chapter includes, from the literature, the underlying physiology, and what is detected and can be inferred by the follow types of physiological measure: heart, skin, facial thermography; breathing; eye movement; electro-encephalography (EEG); and Functional near Infra-Red Spectroscopy (fNIRS). Based on this review, a decision was made to focus on HRV and EDA as measures of workload in rail signallers.

4.2 Introduction

This chapter introduces physiological measures as a potential way to assess cognitive tasks, MWL, and effort, in safety critical industries such as rail. The challenge in rail is, as automation increases, the signaller role becomes a more cognitive monitoring task, less physical task, making the effort required more difficult to observe and measure. A concern is the impact of new technologies, including automation, may exceed boundaries in human capability resulting in the overall system failing to achieve an intended performance or safety improvement. There is a need for adapted or develop new measures of mental workload (overload and underload), and effort, to assess the impact of automation and other new technologies. Current measures of mental workload rely on observation, or subjective workload assessments that interrupt the task or are completed after the task. Physiological measures offer an opportunity to collect more objective, continuous data, without interruption. Physiological measures detect a physical aspect of bodily function such as breathing rate or heart rate. When applied to cognitive activity such measures are referred to as psychophysiological measures. Psychophysiological measures in research areas such as Human Computer Interaction (HCI) provide a new way to seek "a sixth sense" of someone's psychological state (both cognitive and affective) (Dirican and Gokturk, 2011) which has previously been hidden or difficult to measure. In this review the term 'physiological' will refer to what measures detect, whilst psychophysiological will refer to what they can infer.

Research using psychophysiological measures is a growing area as contemporary technologies and computing allow collection of continuous physiological data in real time. Physiological measurement devices are reducing in cost, increasing in robustness, and are increasingly portable. This provides new opportunities for studies to be conducted beyond the laboratory. This review is particularly interested in measures being used in applied and dynamic work settings and tasks to determining whether measures suit use in a simulator or live signalling operational environment. Separate to this is the question of acceptance. The proliferation of wearable physiological measures for personal use shows a growth in cultural acceptance and interest. An example is the prevalence of personal fitness trackers that detect physical activity levels. Fitbit, launched in 2009, has sold over 127 million units by 2021 (Larichia, 2022a). Apple has since grown to have the largest market share with wearable shipped devices totalling 162 million units by 2021 (Larichia, 2022b). The implications of acceptance are considered further in other chapters in this thesis. Wearable devices are developing rapidly, for example the global fitness tracker market is projected to grow from £26 billion in 2020 to £83 billion[8] in 2028 (Fortune Business Insights 2021).

This scoping review will focus on presenting what physiological measures currently exist, what they detect, and what they can infer about MWL and

[8] Based on an exchange rate 1 USD = 0.727 GBP, as checked on xe.com 23.10.2021

other factors surrounding cognitive tasks and experience. This includes what the measures can provide about transient changes and dynamic patterns of mental workload and effort during a task. Studies in a simulator or live operational environment are of particular interest. The review will begin with a brief introduction to the physiology of physiological measures and introduce the concepts of stress and anticipation. A range of psychophysiological measures will then be presented, followed by a short summary of lessons learnt about practicalities. The discussion and conclusion will consider whether these measures detect MWL, what they do detect and infer, and the implications of the findings, including practicalities, for this research. Whilst the research focuses on signallers, the implications could apply to train drivers and other industries facing similar changes to staff tasks with new technologies.

4.3 Method

The focus of this scoping review was to identify and map the types of physiological measures and data used to assess MWL, clarify key concepts and definitions, and gaps (Colquhoun et al., 2014) relating to application in the railway industry or other transportation industries. Limited research was found assessing MWL in the rail industry using physiological data. To reflect the industry application of this research, other industries were considered including Air Traffic Control (ATC), pilots, nuclear, maritime, automotive.

4.4 Underlying physiology

A basic understanding of the body's physiology helps the understanding of physiological measures, and what they can infer of psychophysiology.

To function, the body strives to maintain a state of internal stability known as homeostasis (Sherwood 2013). This ensures the body's internal environment remains steady for cells to remain alive, such as maintaining a body temperature of 37^{0}C (Tortora and Derrickson 2007). The two parts of the autonomic nervous system, sympathetic and parasympathetic *(*Sherwood 2013), maintain homeostasis through exerting opposing influences on the body (see Table 4-1).

Table 4-1 Autonomic Nervous System

	Sympathetic	Parasympathetic
Function	Prepares body for strenuous physical activity “fight or flight” emergencies	Maintains bodily resources. Dominates in quiet relaxed situations
Heart	Increases heart rate and blood flow	Decreases heart rate and blood flow
Skin	Increases sweat across body, including palms	Increase sweat in armpits and groin only
Eye	Dilates pupils	Constricts pupils
Brain	Increases alertness	None

The autonomic nervous system exerts its control on the body in two ways. The first is rapid, through direct electrical signals sent via the nervous system. The second is slower, through glands releasing hormones into the blood stream. The hormones take longer to take effect, but their effect may last longer. This research focuses on the rapid nervous system responses, although the difference in speed of signal and the duration of effect are important principles to consider when interpreting physiological data. This time difference is also important to note when detecting a physiological change in a different part of the body. An example of this is if detecting a change in blood flow in the head, a 6-9 second delay should be expected from an external stimulus (Bunce et al., 2006).

4.5 Stress and anticipation

Stress is a commonly used term, referring to a state of mental or emotional strain or tension. In physiological terms stress is broken down into elements.

A *stressor* is anything that throws the body out of homeostatic balance (Sapolsky 1994) e.g. physical danger, toxins, or strong emotional reactions. If a stressor causes a reaction that is greater than homeostatic mechanisms can counteract this leads to a *stress response* (Tortora et al). This stress response may be experienced as *strain*. The stress response has two parts, firstly the rapid 'fight or flight' response, produced through nervous control, prepares the body for physical activity e.g. running from a lion. Then a slower 'resistance reaction', produced by the release of hormones (including cortisol), supports survival e.g. keep running even when the body starts to tire or is injured. Once a stressor is removed, signals sent via the vagus nerve restore homeostasis or 'tone' (Selkurt 1976). Animals, humans included, are superbly adapted for dealing with short term physical stressors e.g. running from the lion (Sapolsky 1994). Our modern lives can expose us to stressors that are emotional or chronic pressures that make us sick. The implication of this for understanding 'what does success look like' is that a sustainable physiological state may include short peaks of sympathetic response due to strain, but prolonged heightened levels are less sustainable.

Anticipation commonly refers to the expectation or prediction of a future event. It is a way for the mind and body to prepare for a potential stressor. If it is detected in physiological data, this will be chronologically before the external events it relates to. Secondly, there is a possibility that if anticipation is accurate, an individual will not experience strain. This possibility was identified in sports research proposing that if (physical) fatigue is experienced at an anticipated level then the individual theoretically will not be aware of the sensation of fatigue (Swart et al., 2012, and Tucker 2009). The example provided was of a marathon runner will not report feeling fatigue later in a race if they feel how they anticipated feeling at that point and feel able to complete the race. This is despite being measurably physically fatigued at that point in a race. In contrast, a short distance runner in a long race would experience and report fatigue as the distance exceeds their anticipated exertion for a race. This presents potential parallels to the anticipation of cognitive tasks and MWL of signallers. A signaller handling many trains may not report high workload if that number is what they anticipated. The individual experience of an event is linked to the extent to which it matches the experience as anticipated by the individual.

4.6 Criteria for measures for use in live operations

A set of criteria were developed, informed by the domain familiarisation previously completed by the author (see Section 1.6) and by ergonomists and operational signalling staff working at Network Rail involved in preparing for the planned live trial.

(1) Staff can walk around if needed - to talk to colleagues, the shift manager, and reach the telephone at the end of their workstation.

(2) Measures must cause minimal or no task interference, to ensure staff can complete all aspects of their task safety and effectively. This is imperative in a live operational environment where staff are responsible for safety critical tasks and operational performance. This includes allowing all staff to wear prescription glasses if required.

(3) Staff control over data collection. The devices need to be swift to put on/remove for the convenience of staff who have a long and busy shift. Secondly, to assist with staff acceptance, devices would benefit from allowing individual staff to turn on and off data collection with no assistance from a researcher. This would apply to toilet breaks, and also any situation that develops at the workstation that the individual does not want recording. The researcher felt that increasing advocacy, by providing this level of control over data to staff, would be an important aspect of applying the measures in the rail industry.

(4) Sensitive to MWL on own. This relates to the data from an individual measure showing sensitivity to MWL. This would assist in keeping the number of devices to a minimum, for practical purposes, in the field.

4.7 Physiological measures

Physiological measures detect a physical aspect of bodily function such as breathing rate or the electrical signals that instruct the heart to beat. When applied to the study of cognitive activity these are referred to as psychophysiological measures. It is worth therefore distinguishing between what measures detect and what can be inferred from this. The assumption is these measures can infer mental states that are not observable in overt behaviour or verbal report (Hugdahl, 1995). Psychophysiological studies seek correlations of behaviour, they do not claim causal links (Lehrer et al., 2010).

The benefits of applying psychophysiological measures are they are less intrusive to the task than secondary task measures or self-report measures. Secondly the measures can distinguish between short duration events (phasic characteristics) as well as tonic changes over time. Thirdly they may detect relevant data not available through current measurement techniques (e.g. as the individual is not consciously aware of them, or one that is difficult to provide a verbal description for).

This field of research is evolving quickly (Charles and Nixon, 2019) from initially studying the ANS and arousal, to the interaction of complex cognitive and emotional processes (Hugdahl, 1995). Charles and Nixon (2019) provide a literature review of current research specifically on mental workload using physiological measures. The research here considers mental workload, but also overall effort in terms of what is sustainable for a human to maintain whilst completing a task. New psychophysiological studies may in turn endorse or question aspects of existing theories of mental workload and related concepts.

In this section physiological measures concentrate on five categories, reflecting the five areas the autonomic nervous system control shown in Table 4-1. The results are also summarised in a conference poster (see Appendix B). Here, each section explains what the measure detects, how it is detected, and what studies have found it can infer. The order is approximately chronological, starting with older, developed and ending with novel methods:

- Heart
- Skin
- Breathing
- Eye
- Brain

Following this, lessons learnt regarding practicality and ethics will be presented.

4.7.1 Heart rate

Heart related psychophysiological measures returned the largest set of results in the literature. Heart Rate (HR) is the number of heart beats per minutes, detectable at various positions on the body (e.g. fingers, wrist, torso). HR is sensitive to changes in overall MWL in a nuclear control room study (Bainbridge, 1983), task demand in a pilot flight study (Gao et al., 2013) and task duration in a car driving study (Hankins and Wilson, 1998). In a car driving simulator study, involving 15 participants, HR and Blood Pressure (BP) indicated short lasting increases in task demand (Richter et al., 1998). A study of 25 Air Traffic Controllers in the field found that Instantaneous HR (the interval between beats) correlated with number of aircraft, task load, NASA-TLX, and skin conductance (Averty et al., 2003).

HR is so sensitive to participant’s physical movement, and emotions, it is deemed unreliable as a standalone measure of MWL. HR is viewed instead as a useful addition to self-report measures. Heart Rate Variability, which is described next, is deemed in the literature to be a more reliable way to infer MWL.

4.7.2 Heart Rate Variability

Heart Rate Variability (HRV) relating to the electrical activity of the heart as detected by Electrocardiography (ECG). The resulting ECG pattern is described as the R wave, with a distinctive peak at the start and end of each beat cycle (see Figure 4-1). Heart Rate Variability (HRV) measures the varying time gap between these peaks, the R-R interval, and is an important psychophysiological variable (Lehrer et al., 2010). Medical grade measures include multiple leads with sensors placed on the skin in up to 12 positions around the body. In this research a recording sufficient for research purposes can be detected with two sensors in contact with the skin.

Figure 4-1 R wave peaks in Electrocardiography data

HRV correlates with sensory inputs (Riganello, Garbarino and Sannita, 2012). HR increases and HRV reduces with increased anxiety (Riganello et al., 2012). A study of 7 pilots in a simulator found HRV reduced when task demand was high (Lehrer et al., 2010). In a live train with six trains driver participants, Song *et al.* (2014) found HRV decreased before and after a train stop and at tunnels. They reported this as inferring a level of alertness and mental ‘tension’. As an alternative, Nickel and Nachreiner (2003) propose HRV does not indicate MWL, but instead indicates time pressure and emotional strain. These studies demonstrate the varied results found when using physiological data to measures MWL. This is encouraging as it sows potential whilst leaving a research gap to explore the reasons for change. They could all be representing an element of truth by detecting task demand and influencing factors including anxiety, alertness, and emotional strain. Song *et al.* provide a methodological approach that is viable to collect data in live operations applying a data driven approach to identify what points on a journey HRV changed and the potential reasons for this.

Research suggests the frequency of signal may also provide useful information. High Frequency increased when individuals were relaxed and decrease during stress or physical exercise, low frequency was low during high mental tasks (Nagano 2002) and increased with boredom (Schellekens et al., 2000). Johnsen *et al.* (2003) noted that neither HR or HRV returned to base line during the recovery period. This delay or lack of return to base line is a feature of other psychophysiological measures. It is unclear why this is the case. It could be the genuine time needed for the parasympathetic nervous system to reduce the heart rate. It could be a gradual drift in measurement. The literature recommends considering recovery rate as the time to return from the peak value to 50% of the value of the baseline (Hugdahl, 1995). To achieve this baseline values can be recorded prior to a trial starting so that the relative change in values are assessed rather than the absolute values. Absolute and relative values have both been used in research, ideally authors state clearly which they have used.

One ethical consideration is the R spike is the main feature of arrhythmia detection (Kim et al. 2022). Devices are not medical grade, researchers are not medically trained, and such feature detection is not the focus of cognitive studies. Even so, practical and ethical considerations need to be addressed. Secondly wireless ECG devices are not recommended for participants with pacemakers in case they interfere with the pacemaker. To address these issues, research suggests the frequency of signal may also provide useful information. High Frequency increased when individuals were relaxed and decrease during stress or physical exercise, low frequency was low during high mental tasks (Nagano 2002) and increased with boredom (Schellekens et al., 2000).

4.7.3 Skin – Electrodermal Activity

Electrodermal Activity (EDA), also referred to as Galvanic Skin Response or Skin Conductance Response (SCR), is a measure of the electrical conductance of the skin (Venables and Christie, 1980). Sensors are placed in contact with the skin, e.g. wrist or finger, making it easy to apply. Conductance increases as the skin sweats. An individual's state can then be inferred from the level of sweat. To understand EDA data the underlying nervous control by the sympathetic nervous system should be taken into account (Kim et al., 2015).

EDA is good for inferring stress (Healey and Picard, 2005) and arousal. An aroused state describes being awake, alert, and attentive. The term arousal is used interchangeably with 'activation' in the psychophysiological literature (Hugdahl 1995), describing a level of cortical, behavioural, or autonomic activity. The common usage of arousal can imply sexual arousal, here it refers to alertness and autonomic state.

In a car driving study 33 participants completed a driving and braking task in a simulator. Electrodermal Response (EDR) distinguish between different tasks (Healey and Picard, 2005). In a second study of 108 car drivers, HR and Skin Conductance Level (SCL) both increased statistically significantly with changes in cognitive load (Mehler et al., 2012). The study compared different ages groups of drivers and found SCL was lower in 40s and 60s age, but the pattern of incremental increase was consistent across all age groups. This finding suggests varying cognitive load can be inferred from EDA, and that the use of relative values is beneficial for showing patterns in data. Absolute values may be of use if comparing between participants.

In train driving studies, EDR distinguished different train driving and braking tasks, including emergency braking (Collet et al., 2014). A live trial with train drivers trial in Korea found EDA reflected arousal status of drivers (Song et al., 2014). GSR correlated with low adhesion conditions and anticipation of such conditions (Crowley and Balfe, 2018). Drivers were aware of an area prone to low adhesion and GSR spiked in advance of the location. This anticipatory element of EDA offers an interesting additional potential not mentioned with other measures.

A live trial involving ten rail signaller EDA was a good discriminator between high and low self-assessed workload, in two out of three scenarios (Broekhoven, 2016). It was theorised the lack of EDA in the third case was due to a lack of an underlying stress response. The duration of raised EDA levels varied after phone calls were received by the signaller. A short recovery time was associated with a phone call received with factual information and minimum service disruption. A longer recovery time was associated with a phone call where a driver reported they may have hit a person. This call contained levels of emotionality, and uncertainty including potentially longer disruption to train services. The situation was resolved when it was confirmed no person had been hit and services could resume. This study was the only one identified in the scoping review focusing on railway signallers.
Whilst EDA can be used on its own, some authors recommend combining it with other measures (e.g. cardiac activity) to enhance reliability of results (Collet et al., 2014). This is a recommendation made in the literature for many physiological measures.

4.7.4 Facial Thermography

Facial Thermography is a newer psychophysiological measure that detects changes in blood flow to the skin derived from temperature detected by a thermal camera. A laboratory study found that facial thermography and pupil diameter were good indicators of workload, corelating strongly with normalised Instantaneous Self-Assessment (ISA) ratings (Marinescu, et al.,2018). A recent flight simulation study suggested it shows promise as a MWL measure, with the nose area skin temperature correlating highly with both subjective workload measures and performance (Collet et al., 2014). The study did note individual differences, as discovered by other psychophysiological studies in this scoping review. This method is not yet common in the literature but could prove a useful addition to the choice of measures in future. Machine learning techniques could improve the data analysis burden of this measure. Machine learning could be taught to recognise facial features, then use this to more rapidly process the thermal video data that has previously been a more manual processing step.

4.7.5 Breathing

Breathing rate is a measure of how many breaths are taken in a given period of time. Changes in breathing rate when sedentary can infer cognitive and emotional states (Song et al., 2014; Harmat et al., 2015; Rendon-Velez et al.,2016). Breathing rate, and depth, can be detected with a chest strap in contact with the skin. Wireless options are available allowing free movement. The prevalence of such devices has greatly increased in sport science and personal fitness tracking (Sanders et al., 2016). The measurement of breathing is more prominent in physiological studies than in psychophysiological studies.

In the scoping review of psychophysiological studies breathing rate was not measured in isolation. Studies instead combined breathing rate with other psychophysiological measures. A study involving 77 participants playing the electronic game TETRIS found effortless attention correlated with deeper breathing. However, no association was found with blood flow in the prefrontal cortex (Marinescu et al., 2016). In an Air Traffic Control (ATC) study of workload (Harmat et al., 2015) breathing rate, along with other physiological measures, correlated with changes in workload. Breathing rate does not necessarily inter-correlate with other physiological measures.

In terms of practicality, wireless devices reliant on Bluetooth are not recommended for participants with pacemakers as with the ECG device. This can be covered in consent form and participant briefings.

4.7.6 Eye Measures

Eye Movement Tracking (EMT) is a measure of the location of gaze. MWL can be inferred from eye measures including blink rate, fixation time, pupil dilation and eye lid closure. These can be measured either through devices worn on the head, or data captured by remote camera. In terms of detecting MWL, in a driving study, blink rate decreased at complicated sections of road (Brookings et al., 1996). Blink rate was sensitive to short term temporal changes of workload, both in differences in overall task complexity and changes in peak complexity in a nuclear control room task (Richter et al., 1998). Blink rate increased in a monotonous car driving condition (Borghini et al., 2012). In other eye measures, increased MWL correlated with pupil diameter and fixation time but not saccade distance or speed (saccade being the rapid movement of the eye between fixation points) (Gao et al., 2013). Brookhuis and de Waard (2010) suggest Percentage of Eyelid Closure (PERCLOS) is a useful psychophysiological measure of MWL, however (de Greef et al., 2009) found EEG and Electrooculography (EOG), to detect fixation duration, more accurately.

In terms of detecting drowsiness, a car driving simulator study included measures of blink rate and eye closure (Nguyen et al., 2017). The study found slow eye movements indicated drowsiness and changes in beta band brain waves, and a sharp increase in oxy-haemoglobin (detected by Near Infrared Spectroscopy), happened several seconds before eye closure. This study shows how different physiological measures can detect different changes that occur in sequence, rather than simultaneously.

On a practical level eye measurements suit simulator conditions as ambient light is controlled, as data can be affected by ambient lighting levels (Matthews et al., 2014; Tang et al., 2016). A visual task seems the most appropriate and, if using a camera, one where the participant can remain in one position. If a headset is required as part of measurement, duration of study should be considered for the comfort of participants. The difficulty of data analysis is not clear from the literature, but a level of understanding of the underlying psychological mechanisms is implied.

4.7.7 Brain – Electroencephalography

Electroencephalography (EEG) is a non-invasive way of measuring electrical activity in the brain from signals detected on the scalp by multiple electrodes. These electrodes can be fitted ‘wet’ to the scalp by a gel or glue, or ‘dry’ worn in a head cap. The methods of fitment impact how accurate the data is compared to collecting unwanted noise in the signal. EEG produces various brain wave patterns e.g. Delta waves during sleep, alpha waves when awake, beta waves when alert (Kim et al., 2015, Matthews et al., 2014)). EEG data can infer level of activation of areas of the brain and distinguish different patterns of brain wave.

In the literature various studies were identified involving detecting workload and cognitive performance. A study of 92 participants completing a vigilance task found EEG distinguish between two MWL levels, and therefore could monitor vigilance (Kamzanova, Kustubayeva and Matthews, 2014), particularly using alpha waves. Another study of 80 participants found an increase in EEG correlated with working memory, problem solving, reasoning and integration of information, and correlated with both subjective and objective performance metrics (Kamzanova et al., 2014). (Berka et al., 2007) deem EEG a highly sensitive tool, capable of distinguishing high and low workload (Mühl et al., 2014), particularly when combined with eye related variables to remove eye movement artefacts (Zawiah and Dawal, 2016). An

EEG-based cognitive control behaviour study with 37 air traffic controller indicated EEG could discriminate skill, rule, knowledge cognitive levels (Borghini et al., 2017). The findings from these studies are important for two reasons: 1) they show the potential for EEG to detect various aspects of cognitive performance that are relevant to MWL including MWL level, vigilance and SRK level; 2) the need to remove or account for artefacts.

A car driving study found EEG was sensitive to workload changes (Rasmussen, 1983). When car driving difficulty increased, EEG detected theta band increase in the PFC, and alpha band decrease in the parietal area of the brain. During monotonous parts of the simulation, alpha band 'bursts' were detected which indicated drowsiness and reduced vigilance. These findings demonstrate the complexity of EEG findings. The implication for this research is, if EEG is used, it will be important to understand what areas of the brain are involved, and the different types of brain wave, rather than simply a measure of amount.

In rail, in a study of 15 drivers, EEG showed a significant difference between three train driving conditions: daytime, rainy day; and rainy night (Borghini et al., 2012) although differences were not significant across all channels of EEG sensors. A rail study in Korea was conducted in a live operational environment. The study combined multiple physiological measures. Six drivers wore an EEG head sensor, breathing belt, and sensors on the fingers for EDA (Zawiah and Dawal, 2016). The study considered whether measures could detect, and therefore in future predict, sleepiness from arousal status. EDA, EEG and ECG were found to correlate with arousal and mental 'tension'. In the EEG data, beta waves were higher before and after a train stop and tunnels, and negatively correlated with HRV which decreased. The changes detected were thought to be due to changing levels of alertness and mental 'tension'. In addition to the data findings the study stands out as an example of data being collected using physiological measures in a live railway operational environment. Whilst the application of these measures is expanding such trials are still rare in the literature.

4.7.8 Brain – Functional Near Infra-Red Spectroscopy

Functional Near Infra-Red Spectroscopy (fNIRS) is a relatively new measure in the assessment of MWL, with a 2014 paper celebrating 20 years of progress (Boas et al., 2014). Whilst the least mature measure here, it's potential to visualise brain activity (see Figure 4-2) makes it worthy of inclusion. The fNIRS headband (Figure 4-2) is light to wear and fits to the forehead.

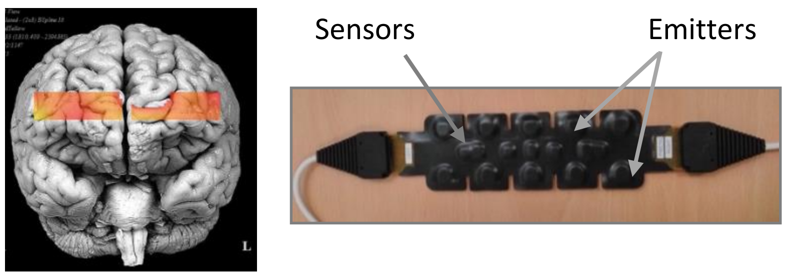


Figure 4-2 fNIRS visualisation of brain activity (right), and headband (left)

fNIRS works by deriving levels of oxygenation and deoxygenation of haemoglobin in the brain, based on the level of near-infrared light absorbed by the cortical tissue. Emitters send near-infrared light signals into the cortex, and the reflected light not absorbed by the haemoglobin is detected by sensors (see Figure 4-2). Based on the signal detected, the level of blood oxygenation and deoxygenation is derived. Cognitive activity is inferred from both levels of oxygenation and deoxygenation. The oxygenation is linked to the expected level of oxygen the brain needs, and deoxygenation indicates how much oxygen has been used.

fNIRS infers cognitive activity in the area of the brain closest to fitment. When on the forehead it is closest to the Pre-Frontal Cortext (PFC), the activity of which is of interest here. The PFC coordinates attention, organises perception over time, goal directed behaviour, and deals with novelty and complexity (Fuster, 2001), and working memory (Song et al., 2014). Humans have higher cognitive capacities in these areas, helped by having the largest and most developed PFC in mammals (Fuster, 2001). Activity in this area of the brain is of interest in this research as signallers are required to sustain attention, and make complex decisions situated in distinct but variable time frames.

In the literature, various laboratory studies found fNIRS was sensitive to MWL. Firstly, fNIRS distinguished three workload levels in an n-back task (Herff et al., 2013), but not in all participants. In a study of air traffic controllers, fNIRS increased with task difficulty in an n-back working memory and attention task (Baddeley, 2001). In a military study of 150 participants completing a threat detection task, fNIRS oxygen saturation was sensitive to workload when compared with base line oxygenation levels (Ayaz et al., 2010). In another military study, fNIRS was sensitive to mental task load and practice level of students completing an unmanned air vehicle task (Ayaz et al., 2012). This would fit with the theory that experts with experience of a task will find it less mentally taxing than someone new to the task. Oxygenation levels were found to have the strongest correlation with performance (Matthews et al., 2014). Findings from all of these studies indicate fNIRS is sensitive to MWL, task difficulty and practice level. Practice level findings fit with theories of working memory effort reducing as experience increases. These studies also raise issues to consider when applying fNIRS including possible individual differences across participants, and the use of relative values (to baseline) rather than absolute values.

The literature also reports on what fNIRS does not detect. fNIRS failed to detect 'effortless' attention whilst participants played the electronic game TETRIS (Maior et al., 2018). In another study, PFC oxygenation was higher with voice rather than visual data communications. The study concluded MWL was lower with visual data. In both cases, these finding could simply indicate PFC did not have a significant role to play rather than providing MWL data. This leads to a broader discussion around whether mental workload only occurs in the prefrontal cortex.

In terms of practicality, to date the literature of studies using fNIRS appears to be focused on laboratory studies. fNIRS is sensitive to motion artefacts which need to be removed during analysis. Using fNIRS beyond the laboratory is in its infancy.

4.8 Practicalities of use: lessons from the literature

There are practical and ethical considerations when using psychophysiological measures. Many can be drawn from the research presented above, with implications for research study design and measures choice. Here is a summary of broader practicalities relevant to any studies using measures to assess highly cognitive, less physically demanding, tasks. The focus here is on the practicalities of applying measures in a live rail signalling environment or signalling simulator.

Firstly, it is common to use psychophysiological measures in combination (Dirican and Gokturk, 2011) or with other measures such as self-assessed workload or observation. This allows triangulation of findings from difference sources, where correlations as well as lack of intercorrelation are of equal interest. Multiple measures can control for confounding variables, such as physical movement artefacts in the data. Physical movement can also be addressed in part through study design requiring participants to remain seated.

Environmental factors such as light and temperature can affect physiological measures so need to be controlled for, especially light levels when using eye measures. This is one of the challenges when moving psychophysiological measures from highly controlled laboratory conditions to a live environment workplace such as railway signalling. Another confounding variable is caffeine consumption, and it has been restricted in some studies (Nguyen 2017) as it is a known stimulant. In this research a decision will need to be made on whether caffeine consumption is restricted in participants or not.

There are medical considerations worthy of note with physiological measures. ECG devices have the potential to detect arrhythmia which**,** whilst not the focus of cognitive studies, raises ethical questions of how to deal with unusual readings. As the researcher is not medically trained a solution is to clarify will all participants that studies cannot be regarded as a medical check.

Psychophysiological measures show dynamic changes in data over time, including duration of effect and proportion of change. This has two practical implications, firstly using relative values (e.g. fNIRS and EDA) may be more appropriate than absolute data values. This means that it is the *change* in value, the direction of change and the timescale of change that are of most interest rather than an absolute value. Secondly, a baseline value is needed to compare to however following spikes in the data values may not return fully to the baseline. Linked to this second point, what would need further investigation is whether a gradual tonic shift over time is a measurement error that should be removed during analysis, or whether tonic changes could be detecting a relevant underlying shift in physiological levels. An example is EDA values will generally shift upwards as the area of skin next the sensor becomes sweaty due to the close-fitting sensor.

Caffeine may be a variable that needs to be controlled for in future psychophysiological studies. Participants in studies can be asked to avoid caffeine (Borghini et al., 2012; Nguyen et al., 2017). This is worth considering for the current research when determining what restrictions, if any, to put on participants.

Another consideration is, of the measures presented so far, EEG appears the most complex in terms of data analysis and correct interpretation of the data.

This has implications for the researcher in terms of the implementation of the measure.

Finally, the ease of data analysis and the learning curve for researcher varies between different psychophysiological measures. This is, in part, due to the extent to which the data collected requires signal amplification or post collection filtering and the extent to which a level of knowledge is needed to interpret the signal. The data analysis challenge also varies with the extent to which measures are very new, or more established (with associated software to assist with analysis). Whilst these considerations are not noted in the literature, they do have practical implications for any researcher wishing to apply psychophysiological measures effectively and successfully.

4.9 Selection of measures for use in live operations

The physiological measures identified in the scoping review were compared with the criteria for use in live operations presented in section 4.6 . As a result of this comparison of measures, wearable measures were determined to be preferable. Results are presented in Table 4-2. The EDA wrist strap was the most suitable. Both Heart Rate and HRV were found to be suitable in three of four criteria. The final decision to focus on HRV were for two reasons. Firstly, that the literature suggested HR itself was not, on its own, sensitive to MWL but that HRV was. In addition, given the PhD research requirement to seek novel contributions to research, HRV was found to be a comparatively recent addition to the physiological measures used for workload. No studies were identified that had previously applied HRV in a signalling setting. The combination of the EDA wrist strap and HRV chest strap were chosen to take forward to the next study.

Table 4-2 Comparison of physiological measures to suit live signalling operations

Criteria / **Measure**	**Staff can walk around if needed**	**Minimal task Interference & can wear glasses**	**Staff control over data collection**	**Sensitive to MWL on own**
EDA (Electrodermal Activity) *Wrist strap*	Good – wireless	Good – Easy to wear	Good – wearer can turn off/ remove	Yes
HRV (Heart Rate Variability) *Chest strap*	Good – wireless	Good – Easy to wear	Medium – wearer can turn off. Fitment in private	Yes
Heart Rate *Wrist or chest strap*	Good – wireless	Good – Easy to wear	Good – wearer can turn off/ remove	Medium – usually used with other measures
Breathing *Chest strap*	Good – wireless	Good – Easy to wear	Medium – wearer can turn off/ remove in private	Medium – usually used with other measures
Facial Thermography *Remote camera*	Medium – must face camera	Medium – no glasses preferable	Poor - researcher stops recording	Yes
Eye Movement Tracking *Remote camera*	Medium – must face camera	Medium – no glasses preferable	Poor - researcher stops recording	Yes
Eye Movement Tracking *Glasses*	Medium - small batter pack	Poor – no glasses	Good – wearer can turn off/ remove	Yes
EEG (Electro-encephalography) *Head cap*	Poor – wired device	Medium – weight on head	Medium – needs assistance	Yes
fNIRS (functional Near Infrared Spectroscopy) *Head cap/strap*	Medium – small batter pack	Medium – weight on head	Medium – needs assistance	Yes

4.10 Discussion

This scoping review has presented what physiological measures exist, whether they detect MWL, and what else they can detect and infer about the effort and experience of completing cognitive tasks. This discussion will consider the implications of these findings for the research, including practicalities, of applying these measures in a live operational environment or simulator.

Physiological measures detect various aspects of the body including electrical signals, electrical conductance, level of reflected light (e.g. fNIRS), and movement (e.g. eye measures, breathing). Such measures can be used to infer cognitive activity, thus gaining the term psychophysiological measures. It is important to distinguish between what measures *detect* and what they can *infer.*

The scoping review found various psychophysiological measures are sensitive to overall MWL. They can also be sensitive to other factors surrounding cognitive tasks and individual experience, some of which sit within MWL and some outside. HR and HRV correlate with overall MWL and task demand, but also to emotional state which traditionally is outside of MWL. HRV reduces with increased task demand and anxiety. Nickel and Nachreine (2003) proposes that HRV does not indicate MWL, but instead time pressure and emotional strain. EDA varies with the cognitive load of difficult tasks, strain (stress) and alertness (arousal). Breathing as a measure on its own does not detect MWL but can be of use in combination with other measures, and one study linked it with 'effortless attention'. Eye measures can detect drowsiness and distinguishes complex tasks (blink rate reduces) from monotonous tasks (blink rate increases). EEG can infer vigilance, sleep/awake state, performance, and alertness from different types of brain wave pattern and location of brain activity. All these measures are sensitive to something, but it is questionable whether they are all detecting MWL. The consensus in much of the literature is to use measures in combination. The extent to which measures inter-correlate varies. If measures do detect MWL and other factors it would explain why they can vary independently. Whilst these measures may not single out MWL, they do offer continuous data in real time to explore the more dynamic aspects of cognitive tasks. They are therefore chosen for use in this research.

The first implication of these findings for the research is to choose measures based on what aspects of MWL, or other factors, are the most relevant. The signalling task requires sustained attention, planning, monitoring (visual and auditory), alertness, working under time pressure and varying MWL between low (e.g. infrequent trains, running to timetable) and high (e.g. late running, perturbations, increased phone calls). Based on these task requirements, suitable physiological measures include HRV, EDA, fNIRS, breathing rate (if combined with other measures), EEG, EMT, Facial Thermography. These measures have been found to correlate with MWL. They can, in addition, show dynamic patterns and changes over time, both phasic (short duration), and tonic (longer duration). The proportion of change and speed of change are useful forms of physiological data. Such data can uncover the physiological effort being put in to achieve a performance outcome. In future this could lead to a better understanding of how sustainable such effort is and indicate whether the task is leading to strain.

EDA appears to detect anticipation of a potential near future event. This anticipatory element was discovered during the scoping review and is of interest to this research. It is hypothesised that novices will not show an anticipatory spike prior to events, but experts will. In contrast, a novice is more likely to show a spike after an event that indicates their level of surprise or habituation. Spikes will reduce after repeated exposures these, a novelty is replaced with familiarity. Whilst these are outside traditional MWL, how predictable or unpredictable events are is relevant to signallers. The signalling task includes dealing with highly predictable timetabled events as well as unexpected disruptions. EDA could, potentially, distinguish between difficult task events that can be anticipated by experts, and events that are unexpected. This becomes less about the *quantity* of task load and more about the difference between *expectation* and *reality*. If signallers expect something to occur, they will show a small physiological change and swiftly return to baseline, reflecting a sustainable physiological state. Unexpected events would show elevated responses in physiological data that take time to return to baseline.

In terms of setting, the literature shows psychophysiological measures suit application in simulator and live environment settings. To date, however, little has been done in rail. This provides an opportunity appropriate to a PhD to apply these measures in a new setting. To successfully apply these measures 'in the wild' however, it is important to control for the confounding variables physical movement, light levels (affects eye measures), and caffeine consumption. In a signalling simulator lighting can be controlled, and physical body movement is mainly arms and head. In a modern Integrated Electronic Control Centre (IECC) signallers are seated (controlling for physical movement) in a light and temperature-controlled room. Signallers however must be able to move along a workstation and around the room walk. This favours wearable measures where physical movement can be recorded by accelerometers and dealt with in data analysis. With regards to caffeine consumption, observation of signallers during the domain familiarisation for this research suggests some signallers consume caffeine frequently during shifts. It may not be appropriate to request signallers abstain from caffeine. Instead, consumption will be recorded, including during data collection.

Regarding validity, the variation in application of measures, and definitions used, make it difficult to directly compare or combine findings from different studies. This leaves a mixed picture of whether specific physiological data are sensitive to specific workload factors. Where correlations exist in the research this indicates a relationship exists but does not prove a causal relationship. Despite this lack of clarity, physiological data remains of interest to this research as there is an interest in the full range of MWL, and understanding what effort underlies sustainable workload that supports successful operational performance. To progress the use of physiological data, an accuracy of terminology is needed in future research to both label the construct of MWL being examined and the specific type of physiological data being measured. This will confirm what certain physiological data are diagnostic of, factors such as task demand, time pressure, stress, or alertness.

This in turn can inform further discussion on tightening the definition of the construct of MWL and the factors that influence it.

Regarding reliability, the evidence base for psychophysiological measures is currently incomplete. No studies in this review collected data from the same participant on occasions separated by days or weeks. Some studies find inconsistency in results between participants (e.g. fNIRS). This second finding could be a genuine difference between participants in their experience of the task. This limited reliability data remains a weakness surrounding psychophysiological measures. To address these limitations in this research firstly a within-subjects study design will be applied, and secondly multiple measures will be applied to allow triangulation of results. Further research, beyond the scope of this research, is needed to prove reliability. Ideally such studies will provide open access data, and report clear data processing methodologies, to build an evidence base and determine the boundaries of appropriate use of each measure.

In terms of the dynamic changes over time psychophysiological data show, the return to baseline can be relatively slow, or a value may fail to return to a previous baseline. After each short phasic event, the underlying tonic value can shift. It is unclear in the literature why values fail to return to baseline. These shifts in tonic values over time may be measurement errors (e.g. EDA device causes skin under the device to become sweatier). Alternatively, measures are detecting a genuine cumulative effect. In EDA this is dealt with by treating recovery as a 50% return towards baseline. If this is a genuine shift in baseline, it may make visible a reducing ‘spare capacity’ in real time, whether or not individuals are consciously aware of this reduction. This area of psychophysiological measures is worthy of investigation and consideration in future studies. The practical implications for this for the current research are that it is important to firstly determine what is an appropriate baseline is for each measure and collect it. HR or HRV can be compared to a baseline value taken at the beginning of a session, whereas fNIRS values need to be compared over a shorter time period. Secondly be able to distinguish between any phasic changes and tonic changes. This can generally be done by looking at the pattern and speed of change (over seconds or minutes). Thirdly, then compare relative data rather than use absolute values of data. This allows for analysis of changes over time, direction of change, and proportion of change, with speed of return to baseline as an indicator of 'recovery'. Finally, seek evidence of a cumulative effect over time that may indicate a shifting quantity of 'spare capacity'.

In terms of feasibility of applying these measures in rail, one key aspect is whether signallers will accept the use of these measures. One key benefit of these measures is they are not intrusive to the task as they require no interruption to the task, as current self-assessment measures require. Acceptance, however, goes beyond whether a measure interrupts the task. Potential distraction is another potential concern. A feasibility study will be conducted prior to trials with signallers to assist in confirm whether the devices may distract from the signalling task or be uncomfortable or impractical (e.g. due to time required for set up). Anonymity of findings will be another factor. Finally, how signallers perceive this type of personal data being used in future is important to consider. These feasibility concerns will fed into the ethics application, study design, and recruitment for studies involving railway signallers.

4.11 Conclusion

To conclude, various psychophysiological measures were identified as linked to MWL. The use of psychophysiological measures is expanding in the literature, including in simulator and live environment settings. The application in the railway industry or in live environments is, to date, small. The measures offer potential in rail, and whilst their novelty in this application suits the need for PhD research to contribute something novel, novelty in rail could add to the challenge of gaining acceptance of signallers in studies.

The measures offer the potential to infer alertness, time pressure, strain (the experience of stress), and anticipation. Ideally measures are used in combination to enable triangulation of results and to assist with controlling for confounding variables. A weakness of psychophysiological measures are their limited evidence of reliability and validity. Further research, beyond the scope of this research, is needed to build validity and reliability evidence. In building this evidence, psychophysiological measures could identify what aspects of the task, or other factors, cause spikes in physiological data to determine the extent to which these findings endorse existing MWL theories. This will be the next step in the research, with a feasibility study to assess the suitability of these measures for assessing the signalling task. After assessing the signaller’s task in this research, in future the measures may suit assessment of other rail tasks.

Chapter 5: Study 2 – Using wearable to infer workload during a simulated signalling task

5.1 Chapter overview

This chapter presents the results from Study 2, a simulation study of twenty participants wearing physiological measures to infer MWL during a rail signalling task. Heart Rate Variability (HRV) and Electrodermal Activity (EDA) data were collected and compared to task demand and self-report workload. The study was conducted to provide an initial test of methods, equipment, data processing and data visualisation prior to future live trialling of measures with industry. Results showed what aspects of workload different measures were sensitive to, with storyboards graphing dynamic changes over time to show the complexity of relationships between HRV, EDA, task demand and subjective workload.

5.2 Introduction

Additional Mental Workload (MWL) measures would benefit railway signalling control as signalling moves from manual signal boxes to larger centralised centres. These centres allow signallers at seated workstations (see Figure 5-1) to use new assistive technologies to control larger areas. At workstations (compared to a traditional signal box) the signallers' task is more cognitive rather than physical, making workload less overtly observable. The Rail Accident Investigation Branch has recommended improving signaller workload assessment (RAIB, 2020), including the impact of new technologies. Whilst measures are already available that measure task demand and subjective workload, an opportunity exists to build on these through physiological approaches. This study considers whether less observable MWL could be inferred from wearable physiological measures.

New MWL measures need to suit use in live operations, where current measures such as subjective workload scales can be intrusive and too time consuming to apply. Traditionally workload assessments in live operations have used observations, subjective measures after a shift, or the Integrated Workload Scale (IWS) designed for signallers (Pickup et al., 2005b) with verbal ratings given during a task. Wearable measures could collect continuous physiological data with minimal task interruption (Charles and Nixon, 2019) from which MWL could be inferred.

MWL is a construct that encapsulates task demand, how individuals experience workload and performance (Sharples, 2019). Signaller tasks include setting routes, operating level crossings, and phone calls with drivers and staff out on track. Signaller workload varies from relative inactivity (PENNA, 2018) to overload, particularly during disruption (Krehl and Balfe, 2014). Figure 5-1 shows Mansfield, an example VDU workstation with information screen to the left, level crossing CCTV screens at the top, track below, and tracker ball and button controls on the desk in front of the signaller, control panels for the level crossings, and keyboard. Signaller performance is essential to staff, passengers, and public safety. Previous

Figure 5-1 Mansfield VDU signalling workstation, EMCC, Derby (Source: Author)

observations show that signallers split their work between monitoring (the screens), planning, intervention (control inputs), communication and quiet time (Balfe et al., 2008, Thomas-Friedrich, 2017). Monitoring, either active or passive, and planning are difficult to observe and measure.

Psychophysiology, which studies the interaction between brain and physiological function, assumes physiological responses mirror cognitive or emotional responses to events, thus providing ‘a “window” into the mind’ (Hugdahl, 1995). Physiological state is controlled by the sympathetic and parasympathetic parts of the autonomic nervous system: changes in physical state, in the absence of movement, imply changes in mental state. Different physiological measures are sensitive to different aspects of MWL (Matthews et al., (2014), Charles and Nixon (2019)).

In this study Heart Rate Variability (HRV) and Electrodermal Activity (EDA) were measured to indicate MWL. HRV is the variation in time between heartbeats in milliseconds, detected from the heart’s electrical impulse (see Figure 5-2), with variation from baseline typically 0 – 40ms Standard Deviation of Normal-to-Normal beats (SDNN). HRV reflects activity in both the sympathetic and parasympathetic nervous system (Tortora and Derrickson, 2007).

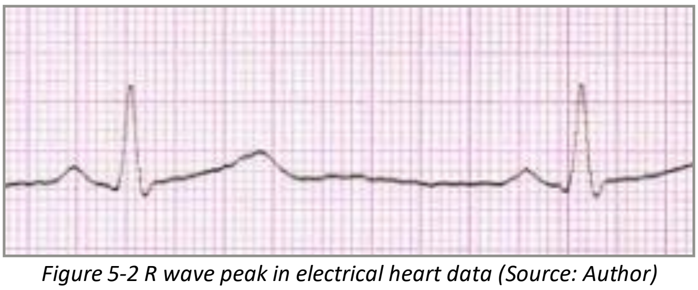


Figure 5-2 R wave peak in electrical heart data (Source: Author)

Research suggests low HRV indicates uncertainty or anxiety (Ramírez et al., 2015), attention (Forte et al., 2019, Colzato and Steenbergen, 2017), and time pressure or emotional strain (Nickel and Nachreiner, 2003). In a study of pilots HRV was sensitive to task demand (Lehrer et al., 2010), and in train drivers HRV reduced during tunnels and at stops (Song et al., 2014) due to 'mental tension'.

EDA is widely used in psychophysiology research (Cacioppo et al., 2017). EDA reflects only the sympathetic nervous system (Tortora and Derrickson, 2007). Level of arousal (alertness) or stress (Healey and Picard, 2005) may be inferred from EDA data that reflects the level of skin sweating (Venables and Christie, 1980). Skin Conductance Level (SCL), the tonic baseline, is typically between 2 – 20µS (Dawson et al., 2017). SCL implies vigilance, attention (Hugdahl, 1995), cognitive activity (Kilpatrick, 1972) or anticipation (Lacey et al., 1963).

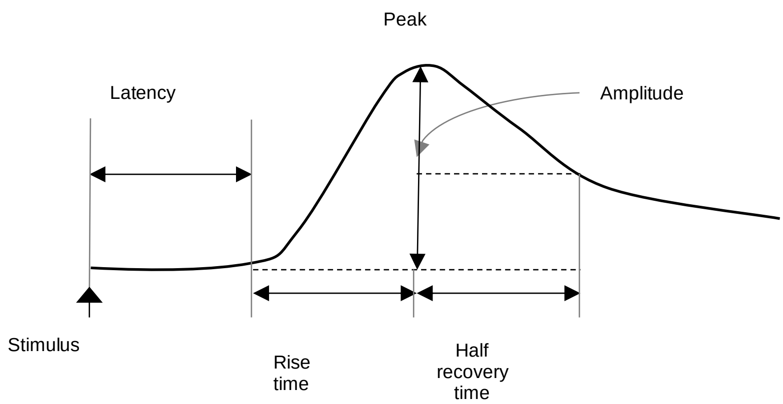


Figure 5-3 EDA SCR response (Boucsein, 2012)

Skin Conductance Responses (SCRs) are peak amplitudes of 0.2–1.0µS above SCL. SCRs indicate a response to sudden stimuli, particularly unexpected events (Sokolov, 1963). Figure 5-3 shows an SCR peak. SCR durations vary with laboratory studies stimulus to peak durations of 2-6 seconds (Cacioppo et al., 2017), compared to up to 14 seconds when stimuli were determined retrospectively from peaks in a continuous task (Bound, 2016). Typical values of half recovery time, in laboratory conditions, is 2 – 10 seconds. SCR amplitudes reduce with repeated exposure, due to habituation (Hugdahl, 1995) as a stimulus becomes more familiar. A study of signallers in live operations comparing three incidents found EDA increase when they received communication regarding an incident. They experienced more frequent EDA

SCRs and their EDA SCL was significantly higher during this period of high self-assessed workload. Their subsequent recovery differed depending on the incidents' significance and operational implications (Broekhoven et al., 2016). The longest was high workload for over forty minutes when they received a call from a driver reporting they may have hit a person (Broekhoven et al., 2016). This highlights the potential wide difference in timescales when discussing workload and EDA in laboratory studies compared to live operational conditions. In a train driving study points of anticipation were identified in the EDA data ahead of an area of low track adhesion, which increases safe braking distances (Crowley and Balfe, 2018).

This study is the first to apply both EDA and HRV measures to assess a simulated rail signalling task. This study considers how changes in temporal physiological data can reflect MWL. The study had the following research questions:

- Do physiological data correlate with task demand or subjective workload?
- Does EDA responses reduce with a second phone call event due to habituation?
- What task events are physiological data sensitive to during a continuous signalling task?
- How could physiological data be visualised to support staff debriefs in the live operational environment?

5.3 Methods

This study applied a mixed methods approach to address a research gap in rail, with limited previous research on physiological measures in this domain. A quantitative approach was taken with the first two research questions to determine whether results replicated previous research: whether physiological data was sensitive to task demand (Lehrer et al., 2010) or self-assessed MWL; and whether EDA spiked occur due to phone call communications (Broekhoven 2016) then reduced due to habituation (Hugdahl, 1995). The latter two research questions took an exploratory qualitative approach to explore: 1) what was occurring that could explain the MWL experienced during a continuous task, 2) how to progress the measures for use in a live operational environment. This method was chosen as previous research suggests that a more complex relationship exists between physiological data and MWL, with physiological data being sensitive to divergent aspects of MWL (Matthews et al., 2014, Charles and Nixon, 2019). These include alertness (Healey and Picard, 2005), anxiety (Ramírez et al., 2015), and emotional strain (Nickel and Nachreiner, 2003). Previous studies of MWL with physiological data that have used advanced statistical analysis and spectral analysis have suited the laboratory setting, where conditions are short and workload levels are controlled. Here instead a timeline data presentation is used (Gillis, 2016). The storyboards and qualitative methods of labelling used here are a data driven approach to identify key points during the task and explore the reasons why these changes may have occurred in the data. An analogy is if studying fuel economy in car driving, maximum speed and total journey time does not provide the whole story. Instead, a timeline identifies when on the journey speed was reduced and a storyboard of why that may be (e.g. traffic). The findings from this exploratory qualitative approach could inform future quantitative studies on what the physiological data are sensitive to, and their validity as measures of MWL, to improve the diagnosticity of physiological data.

The final research question addresses progressing the implementation of these measures from the laboratory to the field, this study explored the use of storyboards from chronological data to visualise data changes in a continuous task. This study provided the opportunity to develop and test a methodologically swift method to graph the data. This would be used to produce a timeline to make the complex data accessible to operational railway staff. Developing this method to be swift would enable physiological data to be use in post-task debriefs with staff, to explore what physiological data can tell us about the task demand, task events and workload. The aim, ultimately, is for such a method to be applicable in live operations with staff in the railway industry.

5.3.1 Participants

Twenty participants took part (11 male, 8 female, 1 preferred not to specify), aged 22 – 39 years (M = 30.2, SD = 5.1). All were university students or staff members, incentivised by a prize draw for a £40 shopping voucher. None had prior experience of signalling. The rational for choosing these participants was they were within the age range of railway safety critical staff, and they were an easier population to access to provide an initial test of the methods and equipment prior to future trialling within industry with railway staff.

5.3.2 Simulation Protocol

The signalling task used SimSig, an ecologically valid simulation of signalling screens used in operational IECC signalling centres in Britain. This PC simulation runs in real time, with trains appearing on the screen based on a real timetable (Figure 5-4 and Figure 5-5). The simulation uses donationware and commercial software. Participants each received training in how to operate the sim followed by 20 minutes practice during which they controlled the simulation to confirm they could correctly signal and route a train. Participants then put on the wearable measures (on their non-dominant wrist and on torso). A 20 second baseline (Averty et al., 2003) was recorded, and an audio clap used for retrospective data synchronisation in time across devices. Task time was 47 min – 64 min (M = 55, SD = 4). The study was approved by University of Nottingham's Faculty of Engineering Ethics Committee, as shown in Appendix C.

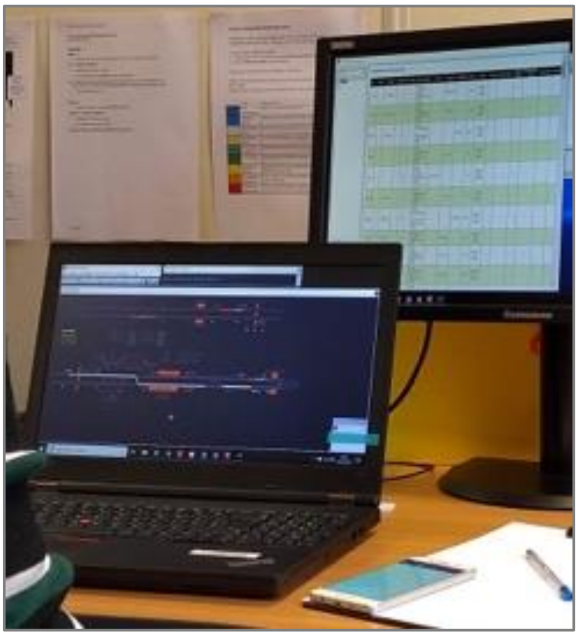


Figure 5-4 Simulation, with timetable and workload app

The number of trains was used as a measure of task demand as each train represented a required set of actions. These can be broadly categorised into planning, intervention (control inputs), monitoring (the screens), and communication (Balfe et al., 2008, Thomas-Friedrich, 2017). During the task participants must notice each train when it arrives on the screen, then identify its destination from the headcode (train reporting number label) by referring to the simplifier (timetable on the second screen, see Figure 5-4).

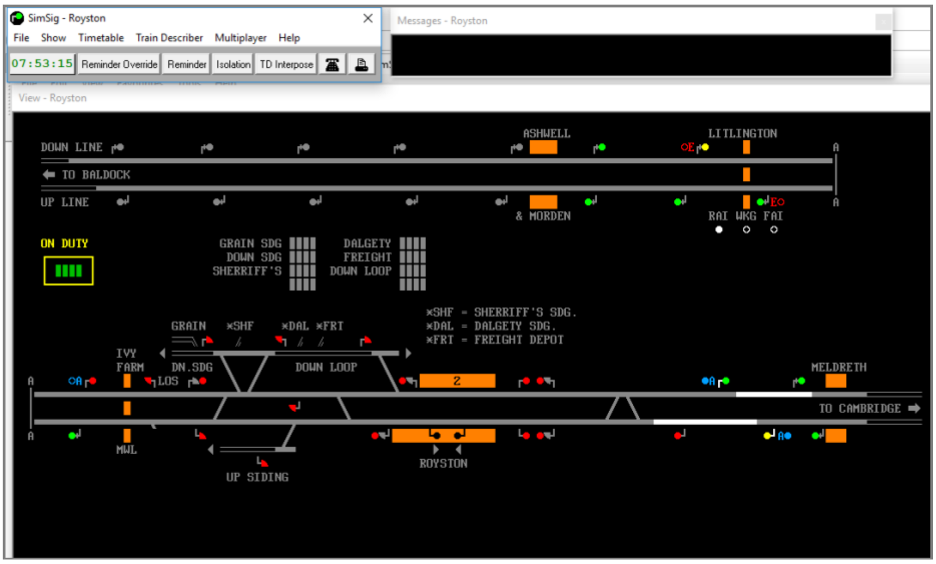


Figure 5-5 Royston, SimSig simulation

They must then set the correct route, the first part only so as not to block the later stages of the route of other trains. Once the route is set the train's progress must be monitored through sections, to notice when the next route can be set. Routes cannot be set in advance if the track ahead is currently occupied. They can only be set when the route shows clear. The number of trains therefore represents all these steps, and associated cognitive workload, to provide a task demand value.

An addition to the design was a phone call. This was included as calls are a potential source of task interference. Whilst communication is a known factor in signaller workload, phone calls are often missing from simulation-based workload assessments, or modelling, of signaller workload (Delamare et al., 2016; Gillis, 2016). Task interference from phone calls was predicted to be one of these critical points as they occurred were an event that was unplanned.

The task was designed to be achievable by individuals with minimal training prior to simulator. The total number of trains was low, relative to busier areas of control. An ODEC workload assessment (Pickup et al., 2010) rates the area controlled as low workload (out of low, medium, high) based on the number of station platforms, junction, depots, controlled signals, and trains. This area provided variation in routing whilst keeping number of junctions and trains sufficiently low that participants could route all trains in the task without extensive training. The timetable had gaps to provide time to complete all tasks including addition tasks such as interposing (changing the train headcode). Task performance was not assessed. This fits with the agreed railway industry future use case of using these measures to assess workload not staff performance. The focus of data collection was on detecting physiological data continuously during a task with varying task demand. It was sufficient for this purpose for participants to demonstrate they engaged with the task. Participants were all deemed to have actively participated in the tasks. If they had not, the number of trains would have increased without a decrease, as trains stopped at red signals awaiting routes.

Task time varied to ensure each participant experienced increases and decreases in task demand, and similar overall total task demand. Task demand was defined as the number of trains on the screen at any one time. A timetable, on the second screen, provided details of when and where to route trains, and when to interpose (change) train headcode (train reporting number label).

To increase the face validity of the task, two phone calls and a freight train were included as randomly timed novel events. A written script was provided for the calls. In the first call, from someone working near the track, the participant listened to the caller's report, noted down their location, confirm they could commence work, and state the time. The second call from the same individual was confirmed work was complete, with the participant noting the time work was complete. Verbal instructions were provided to route the freight train. Table 5-1 presents the planned task times. Actual times depended on the extent to which participants kept to the timetable and correctly routed the trains.

Table 5-1 Planned variation in number of trains and order of events

Time (min)	Number of trains and order of events
0m	0 trains at session start.
1m – 15m	Up to 3 trains, reduces to 1 – 2. Interpose 1 train headcode
At any time	Freight train arrives
15m – 30m	Up to 4 - 5 trains, reduces to 2. Interpose 1 train headcode
30m – 50m	Up to 4 trains, with 2 phone calls
50m – 60m	Session ended after next reduction in number of trains

5.3.3 Equipment

An Affectiva QTM wrist strap (Figure 5-6) recorded skin conductance at 8hz and accelerometer data. A Zephyr BioHarness (Figure 5-7) chest strap

Figure 5-6 Q Sensor for EDA data

Figure 5-7 Zephyr Bioharness for heart data

calculated HRV from a rolling 300 heartbeat SDNN (Standard Deviation of Normal-to-Normal beats) in milliseconds, updated once per second.

A subjective workload measures was included to compare data from the physiological devices with an existing MWL workload using in rail to capture time domain workload ratings. A mobile app, developed by the Mixed Reality Laboratory, University of Nottingham (on desk in Figure 5-4 displaying the Integrated Workload Scale (IWS) designed for signallers (Pickup et al., 2005b) (Figure 5-8). The scale appeared with an audible alert after 2 minutes, remained on screen until a rating was entered, then the screen went blank between ratings. If participants missed ratings, the scale would remain visible until a rating was given. Missing values were omitted from the analysis rather than being estimated. Screen capture software recorded audio and visuals to time stamp individuals' task demand, task events, and all interactions with the task.

5.3.4 Data processing

One participant was excluded due to a technical problem. HRV represents 18 participants as one participant's data failed to record. Missing values were excluded from the study. Mean number of IWS ratings was 21.3 (SD = 5.2), with a total of 405 ratings across 19 participants. EDA data represents 12 participants as

Figure 5-8 IWS scale

seven non-responders were excluded with SCRs <0.05µS (Boucsein, 2012). Data across devices was synchronised by timestamps converted in MatLab.

The data driven investigation of HRV identified task times when HRV was in the lowest or highest quarter for each participant. Within these, each 30 sec period was coded into two types of behaviour:

- Correct (action required and completed, or cursor was still when no action was required)
- Uncertain (interactions were initially incorrect, missed, in error (e.g., wrong route), hesitant, repeated, or cursor moved unnecessarily).

These categories were developed from observations made by trainers of signallers during the domain familiarisation. They described individual trainees who they deemed ready to be passed out (take their test to qualify) were observed by trainers to make correct observable actions. Trainees requiring further training time were observed to make uncertain actions e.g., hesitation. Coding required task knowledge on the part of the researcher to identify which action was required when. Cursor movement provided a clear visual indication from the screen capture.

The data driven investigation of EDA sensitivity identified when EDA peaked. Screen capture and audio data were then reviewed from 30 sec prior to rise times of EDA SCRs and SCLs. This applied the principles of a method used by Bound (2016) to identify events in a continuous task that precede peaks in EDA data. This is achieved by a visual inspection of EDA data, then an inspection of task data in the period preceding the EDA peak. Bound constrained their inspection to 14 seconds prior to peaks. In this study, this period was extended to 30 seconds. This was done to include and incremental increase of SCL due to a series of small SCRs, that lead to a cumulative increase in SCL despite individual SCRs being small. This exploratory data driven approach was applied to inform the inclusion, or not, of this method in analysis of future live operational trials.

Wrist movement was categorised as high, medium, or low. Physical movement is a known confounding variable of physiological data, so those classified as showing high levels of movements were excluded from further analysis. Medium and low were analysed further, to seek stimuli (external task demand or internal response to the task) that could explain EDA peaks beyond solely physical movement.

5.3.5 Statistical analysis

Correlations were analysed using Pearson's correlation coefficient (2 tailed) in SPSS. A *P* – value of < 0.05 was regarded as significant. The difference in observations when HRV was High and Low was analysed using a Paired Samples *t*-Test.

5.4 Results

5.4.1 Storyboards of temporal physiological data

Task demand, task events, EDA, HRV and IWS were graphed across the task in Storyboards for each participant, similar to the time sequence of physiological data of train drivers (Song et al., 2014). Figure 5-9 and Figure 5-10 show examples.

5.4.2 Sensitivity of EDA to workload

A visual inspection of storyboards determined SCRs did not coincide with phone calls, with no associated habituation between Call 1 and Call 2. A data driven exploration of observable behaviour preceding SCRs instead determined that SCRs could be explained by moments in workload including:

1. Moments of realisation
2. Uncertainty
3. Time pressure.

Moments of realisation occurred after an error. Two participants showed an SCR. One saying “Oh oh oh” immediately after dealing with a level crossing (P6, as shown in Figure 5-9 at 33 minutes task time). The second, after realising trains were late said “Disaster. I’m very anxious now, what did I do?”. Two participants showed SCLs on realising an error. One said “Sorry I thought I knew what I was doing” whilst receiving instructions for routes, and the other, later in the task, “You kick yourself, what have I missed?”. These peaks occurred *after* the error, at the time the individual realised their error, such as setting a wrong route.

Uncertainty was observed in participant behaviour around calls, novel events, errors and unexpected system responses. Calls were associated with two participants’ SCRs and one SCL. One participant with an SCR in Call 1 asked “What if I don’t understand [the caller]?” and one with an SCL, higher in Call 2, said after both calls “Sorry I didn’t really catch their location”. In both cases it was not receiving the call that led to the EDA response, but uncertainty and confusion over the content. P06, shown in Figure 5-9, showed a similar response after Call 1 but did not verbalise, so the reason for their SCL is unclear. The novel freight train was linked to SCRs in three participants whilst receiving verbal instructions to route it. P6 showed SCRs after noticing the freight train, during routing it (Figure 5-9). Dealing with the freight train was linked to EDA responses in some participants. Participants received verbal instructions to route it and had to wait for permission to route it into the sidings. Their responses could be due to the novelty of the task, and their associated uncertainty, as the freight route differed to all other services.

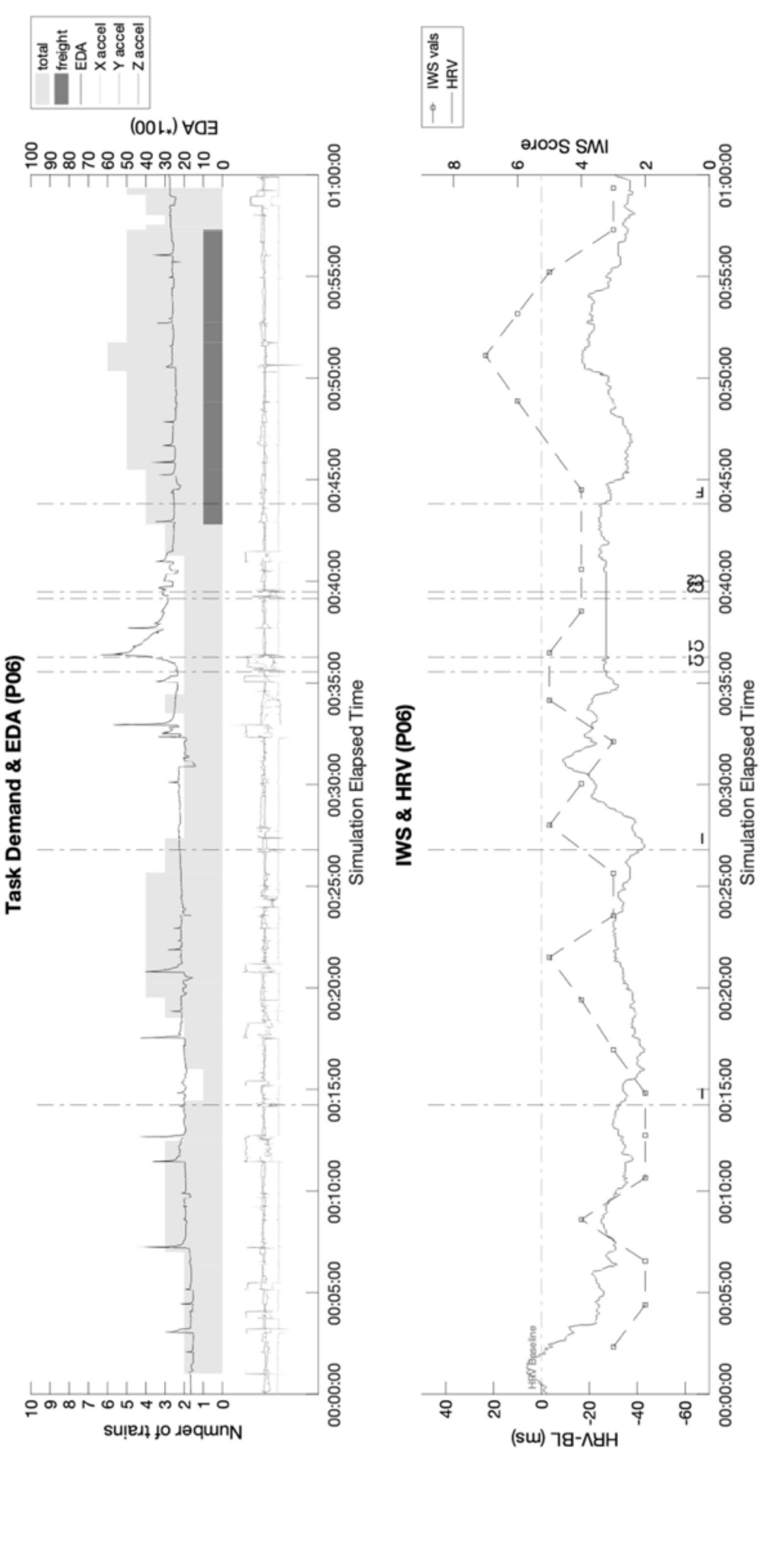


Figure 5-9 Storyboard (P6)

Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).

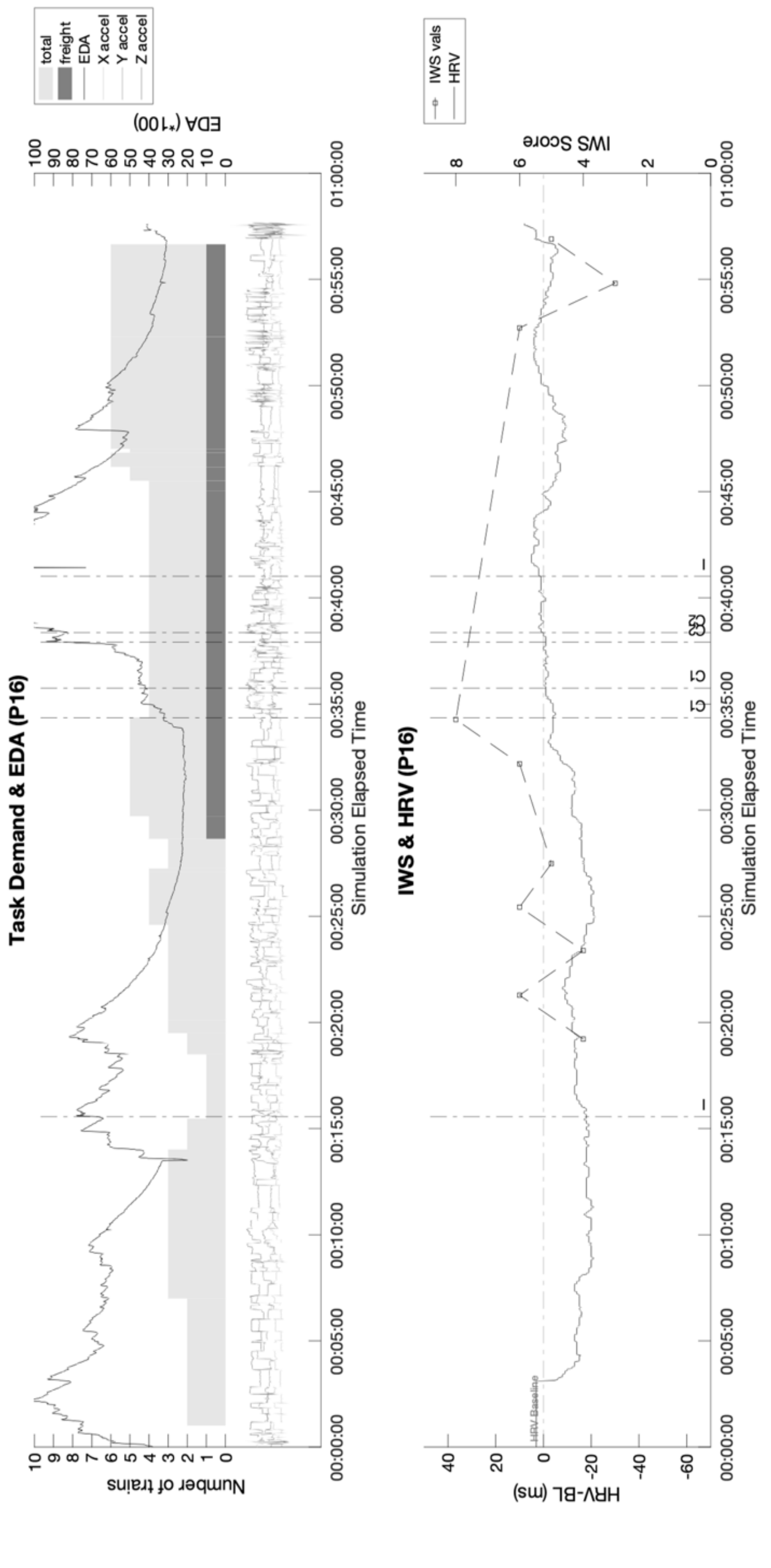


Figure 5-10 Storyboard (P16)

Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).

Further examples of uncertainty included participants unaware of their own errors and confused by unexpected system responses. One participant showed SCRs and one SCLs whilst under instruction to route timetabled trains. Both participants made errors they were unaware of. P16 had a SCL (see Figure 5-10, from 34 minutes task time) when they failed to notice the freight train and set a route into an occupied platform. Their SCL continued to increase through both phone calls and whilst trying to cancel a route. This included a period of failing to answer the IWS. Two participants showed SCRs with unexpected system responses such as not understanding why a train had stopped or a cancelled route taking time to clear.

Time pressure was observed in one of two ways. Firstly, SCRs whilst waiting for an external task to complete, such as a route to clear. Secondly, when participant error resulted in trains running late. In the latter case, time pressure was implied by participants' observed behaviour or comments such as "I'm very anxious now, what did I do? All [the trains are] late!". P16 showed an SCL whilst waiting and trains were late (see Figure 5-10, during EDA rise time 13 – 16 minutes task time).

5.4.3 Sensitivity of HRV to workload

During periods of low HRV proportionally more Uncertain behaviours were observed (e.g., errors or hesitancy) with an associated decrease in Correct behaviours (see Figure 5-11). A Paired Sample *t*-Test (2-tailed) found no significant difference in Correct behaviours between High and Low HRV (M = 7.22, SE = 2.47 df 17, t = 0.21). This suggests a trend towards uncertainty when HRV was low.

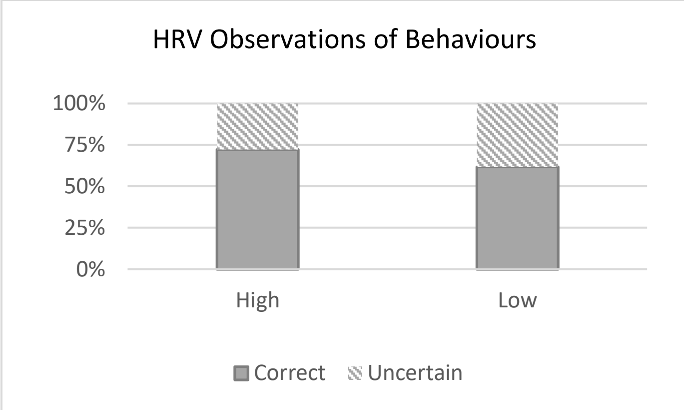


Figure 5-11 HRV Observations of Behaviours

5.4.4 Correlations between HRV, EDA, IWS and task demand

Statistical analysis compared means across the task and values at each IWS rating. The following HRV values were compared with IWS Mean across 18 participants: HRV Baseline with IWS Mean; HRV Mean with IWS Mean; and HRV – BL Mean (HRV Baseline removed) with IWS Mean. Each comprise 366 IWS ratings across 18 participants. Descriptive statistics for these variables are presented in Table 5-2.

Table 5-2 Descriptive statistics for IWS and HRV mean correlations

Variable	Mean (N = 18)	Standard Deviation (N = 18)
Task Demand	3.30	1.11
IWS	4.03	1.64
HRV Baseline	74.08	0.67
HRV Mean (20s to IWS)	64.66	18.35
HRV - BL Mean (20s to IWS)	-11.60	17.26

A strong negative correlation was found between HRV Mean and IWS Mean (r = – 0.721, P 0.001), across 18 participants, 366 IWS ratings. Results of the correlations are presented in Table 5-3. A relationship exists between participants' average subjective rating and their average HRV data. A moderate, significant, correlation was also found between HRV Baseline and IWS Mean (*r* = – 0.570, *P* 0.013). Individuals' HRV Baseline prior to a task provided a moderate prediction of average IWS during the task. This suggests High MWL may be inferred from low HRV. HRV baseline before the task may also partially predict MWL reported during the task. In comparison, when baseline values were removed, HRV-BL Mean did not correlate with IWS Mean (*r* = 0.137, *P* 0.587), see Table 5-3. Removing the baseline corrected for individual differences, leaving workload due to task demand. This finding indicates average IWS had a stronger relationship with individual HRV than with task demand (number of trains).

Table 5-3 IWS and HRV mean correlations

*(18 participants). Bold **significant at 0.01, *significant at 0.05 (both 2-tailed).*

	IWS Mean	HRV Baseline	HRV Mean (20s to IWS)
IWS Mean	1		
HRV Baseline	**= – 0.570* P = 0.013**	1	
HRV Mean (20s to IWS)	**= –0.721** P = 0.001**	**= 0.849** P = < 0.001**	1
HRV - BL Mean (20s to IWS)	= 0.137 P = 0.587	**= – 0.755** P = < 0.001**	= – 0.294 P = 0.236

Two comparisons of HRV and IWS were conducted at the time of each IWS rating, with HRV averaged in the 20 sec period prior to IWS rating: HRV (20s to IWS) with IWS; and HRV – BL (20s to IWS) (HRV Baseline removed) with IWS.

A moderate significant negative correlation with HRV (*r* = – 0.391, *P* < 0.001) was found with individual IWS ratings, see Table 5-4. This fitted with the correlation found between task HRV Mean and IWS mean. When HRV Baseline was removed, to correct for individual differences, the correlation with IWS was absent (*r* = 0.036, *P* < 0.492).

Table 5-4 HRV, EDA and Task Demand correlations at times of IWS ratings

*Bold ** significant at 0.01 level (2-tailed); * significant at 0.05 level (2-tailed). Sample sizes vary due to data processing, and number of IWS rating per participant. N = number of values from IWS of 19 participants, HRV of 18 participants and EDA of 12 participants.*

	Task Demand	IWS Rating	HRV (20s to IWS)	HRV-BL (20s to IWS)	EDA
Task Demand	1				
IWS	**0.474**** **P = < 0.001** **N = 405**	1			
HRV (20s to IWS)	**– 0.154**** ***P* = 0.003** **N = 366**	**– 0.391**** ***P* = < 0.001** **N = 366**	1		
HRV-BL (20s to IWS)	0.002 *P* = 0.962 N = 366	.036 *P* = 0.492 N = 366	**0.169**** ***P* = 0.001** **N = 366**	1	
EDA	0.112 *P* = 0.072 N = 260	**0.294**** ***P* = < 0.001** **N = 260**	**– 0.147*** ***P* = 0.029** **N = 222**	**– 0.158*** ***P* = 0.019** **N = 222**	1

At the time of each IWS rating HRV was a weaker predictor of a specific IWS rating. Figure 5-12 illustrates this difference between absolute HRV and relative HRV – BL values.

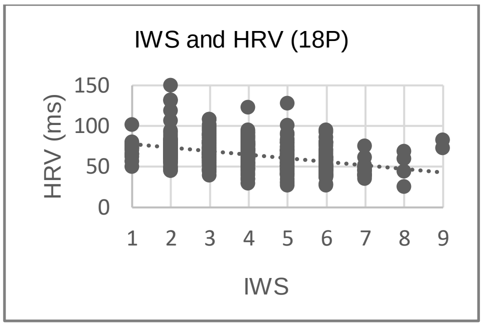
Task demand and IWS showed a moderate significant correlation across 405 IWS ratings by 19 participants (r = 0.474, P = 0.001), see Table 5-4. IWS ratings changed partly in line with task demand. In comparison task demand and HRV showed a small significant negative correlation across 366 ratings by 18 participants (r = – 0.154, P = 0.003), see Table 5-4. Task demand and HRV-BL did not correlate (r = 0.002, P = 0.962). Once individual differences were corrected for, HRV did not predict task demand.

Task demand and EDA did not correlate across the 12 participants showing an EDA response (r = 0.112, P = 0.072). Descriptive statistics are presented for EDA in Table 5-5.

Table 5-5 Descriptive statistics for IWS and EDA mean correlations

Variable	Mean (N = 12)	Standard Deviation (N = 12)
Task Demand	3.30	1.11
IWS	3.97	1.64
EDA	0.20	0.12

EDA did not show a relationship with task demand. As task demand increased HRV decreased, but to a lesser extent than IWS increased.



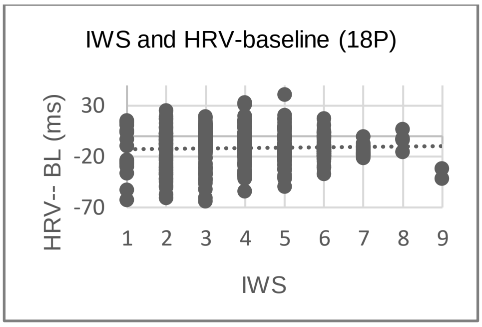


Figure 5-12 IWS with HRV and HRV-BL

IWS and EDA showed a small significant correlation (r = 0.294, P = < 0.001), see Table 5-4, for the 12 participants who were EDA responders. Whilst EDA does not correlate with task demand, increases in EDA does partially infer increased MWL in those participants who record an EDA response.

In summary, HRV showed a negative correlation with IWS and Task demand. HRV showed a stronger association with IWS than with task demand. These HRV correlations came from values that were not baseline corrected. EDA showed a small significant correlation with IWS (in responders), but phone calls did not cause SCRs.

5.5 Discussion

Findings show a complex relationship between Task demand, IWS, HRV and EDA, with responses varying across participants. The temporal data presented in storyboards provide a detailed visualisation of this complex interaction across a dynamic task. They demonstrate the potential to be produced swiftly in a live environment to form part of a debrief with operational staff. This use case could prove valuable to prompt discussion with staff as to why MWL or physiological data changed at certain points in time during the task. This method would add validity to the qualitative labelling of events in the continuous data by involving the individual who performed the task.

A strong negative correlation was found between HRV Mean and IWS Mean across the task, showing convergent validity for mean HRV. This finding matches previous studies that show HRV decreases as workload increases. At individual instances of IWS rating, a moderate negative correlation was detected. Whilst HRV may not be a strong at predictor of specific a subjective workload rating, it could potentially predict the average subjective workload rating across a task. This finding fits with previous research that subjective workload ratings are good indicators of how an individual experiences the workload, rather than reflecting task demand. Findings suggest, if only an average workload rating was required for a whole task, HRV could replace IWS. HRV values, averaged over the task, could provide an estimate of the likely average subjective workload score. In addition, an HRV baseline may be sufficient to provide an indicator of where on a subjective workload scale, for a similar task, individuals will report workload relative to each other. This would fit with HRV indicating individuals' underlying current autonomic state, and where they currently sit in the balance between the sympathetic and parasympathetic nervous system. A sample of HRV may, in future, provide an individual 'weighting' to complement subjective ratings.

The above correlation between HRV and IWS was removed when HRV data was corrected relative to baseline. HRV – BL did not predict IWS ratings. Whilst it is common with physiological measures to use relative values, with baseline removed, this study found HRV absolute values can offer insight into MWL. In this study the same task demand (timetabled trains) led to different individual experiences of workload. Absolute values showed differences between individuals' experience of workload, whilst relative values (baseline corrected) provided data on task demand. Both have their place. In terms of MWL however, the meaning of absolute values in physiological data are not currently well understood.

Task demand (number of trains) and IWS showed a moderate significant correlation. This suggests subjective workload was more than simply the number of trains on the screen. At instances of IWS ratings, HRV and IWS showed a moderately significant negative correlation that was stronger than the HRV and task demand small significant negative correlation. This suggests HRV indicates individual effort and time pressure (as reported by IWS), more than HRV reflects the number of trains on the screen.

In terms of the individual experience of workload, a trend was found for HRV being lower when uncertainty was observed. Whilst no significant difference was found between high and low HRV, uncertainty linked to HRV matches previous research by Ramírez *et al.* 2015. Taken together these findings show HRV can imply individual workload experience.

EDA identified moments in workload during the task. EDA did not correlate with task demand and showed only a small significant correlation with IWS. Instead, both EDA SRCs and SCLs indicated moments of realisation, and periods of uncertainty, or strain due to time pressure. Moments of realisation occurred when an individual realised something such as an error. The timing of this realisation was not when the error was made (when they had no awareness of erring). This finding matches the concept in literature of SCRs occurring due to a sudden unexpected, or novel, event. In this study they were internal sudden realisations, not external stimuli.

Determining when physiological data showed 'stress' or 'alertness' proved challenging. Instead, in this study 'Uncertainty' was used to label observations of behaviour that could then be linked in time to physiological data. In in the EDA data, uncertainty was observed with calls, novel events, errors, and unexpected system responses. These were based on what participants were heard to say, or their speed and accuracy of cursor moves and interactions. The literature terms these responses as level of alertness, stress, or anticipation. There is therefore some ambiguity in the terms used by different studies to describe what may be a similar underlying physiological response. To develop these measures, clarity is needed when reporting results to distinguish what the data *detects* and what it *infers* (e.g. stress, alertness etc).

In the EDA data, SCRs did not coincide with phone calls or show habituation between calls. Similarly, the arrival of the freight train did not result in EDA responses for all participants. Whilst this does not match previous laboratory studies of sudden external events, it does match findings in the field of EDA response varying with the reason for communications (Broekhoven et al., 2016): EDA appeared to indicate only individuals for whom the calls caused confusion.

Whilst a proportion of non-responders is to be expected in a study, this study had a higher proportion of non-responders with only 12 of 19 participants displayed SCRs or SCLs. This could be due to cooler room temperature, or sensor placement. In addition, this result may be a valid reflection of the simulator setting, and a task that some participants found easy to complete without any sudden unexpected events. EDA increases only when the situation is deemed to need more attentional resource, or effort.

Results suggest aspects of experienced MWL can be inferred from physiological data. Physiological data was found to have a stronger relationship with subjective ratings than with task demand (number of trains). This suggests physiological measures strengths suit measuring individuals' experience of workload, and how this varies between participants. In addition, it could indicate that the increase in number of trains did not account for task demand. Time pressure, complexity and Novelty of Events (e.g. dealing with the freight train) are likely to have contributed. EDA's strengths lie in pinpointing moments pertinent to workload in a continuous task. HRV can indicate an individual's underlying autonomic state during a task. Contextual information is required to determine what has led to a physiological response. They both could potentially, in real time, indicate an individual's changing capacity to respond to task demand. Whilst absolute values of physiological data may yet to be fully understood, individual patterns of HRV changes could be of use.

5.5.1 Implications and future work

Physiological data could be used in a debrief, to address aspects of the task that result in the largest physiological responses. Identification of precise of workload, in graphed storyboards of the task, could enable the provision of a retrospective narrative of individual workload. This narrative approach would provide detail that cannot be captured during the task without interruption. Such an approach could be used as a training aid, or workload assessment tool.

This study found individuals differ in their physiological data. To develop physiological measures for MWL assessment it will be important to understand the impact of individual differences. A repeated measures study, with data collected from the same individuals, would assist in understanding how stable (and therefore predictable) individuals' responses are over time.

To progress the use of physiological measures for MWL assessment it is recommended further research distinguishes the findings of what data were *detected* from what was *inferred*. This will assist in clarifying what each measure is best suited to assessing. In this study electrical signals from both skin and heart activity were detected. From these data individuals' experience of workload were inferred. Whilst research continues to decipher the

interaction between physiological data and MWL, this distinction will assist clarifying the diagnosticity of physiological data for MWL assessment.
In addition, these measures cannot be applied in isolation. Where inferences are made, these rely on additional sources of data, such as event logs or individual accounts, to account for changes in data.

This study focused on what could be detected from continuous temporal data, with qualitative analysis exploring what events in task and physiological data co-exist in time or in sequence. This is exploratory approach would benefit in future from the outcome of qualitative debriefs with staff. If clear sequences of events are identified these could then inform future qualitative research with advanced statistics.

Finally, applying wearables in live rail operations is an emergent field. If wearable measures are to use physiological data in future to assess staff MWL, it will be important to understand the perspective and attitudes of staff to both wearing the devices and the use of their physiological data.

5.6 Limitations

Participants had no prior experience of signalling, and practice time on the simulator was limited. Whilst all participants successfully routed trains, some participants seemed to experience uncertainty or confusion when unexpected system responses occurred or made errors that they were initially unaware of. This led to a wider range of experience of the task than only the varying task demand. The resulting varying physiological data did suggest this element of identifying moments of uncertainty is worthy of further investigation in a future study.

The study had 37% non-responders, with SCRs <0.05µS (Boucsein, 2012). This is higher than the 10% of previous studies (Braithwaite et al., 2013). This could be caused by sensor placement or low room temperature (M = 14.9 0C, SD = 2.6). To improve future EDA data collection, sensors could be placed in the volar wrist position or palm, and ambient temperature increased and controlled.

The coding of Correct or Uncertain behaviour required an understanding of the signalling task to determine whether an action was required at any one time. Should future studies apply this method, the reliability and validity of this observational coding would be strengthened by a second observer coding the data to provide inter-rater reliability.

In using HRV data, a longer baseline period may be appropriate in future studies. This would confirm whether the difference between HRV and HRV-BL found in this study endured.

5.7 Conclusion

Additional MWL measures would benefit the rail industry to measure MWL from relative inactivity to overload. This study considered whether physiological wearables could provide a new measure for signallers. The study found that different physiological measures were sensitive to different aspects of MWL.

Both HRV and EDA detected physiological responses that imply, in combination with other data, aspects of individuals' experience of workload. HRV could be used in place of IWS, or other subjective workload subjective measures, particularly if average MWL ratings are required. EDA could be used to identify moments in workload during a continuous task, such as moments of realisation, or periods of uncertainty, or time pressure. It is important to be clear what question needs answering to choose a physiological measure sensitive to detecting an answer or inferring one. The moments in workload could inform a debrief, and focus training efforts, on aspects of the task individuals showed strong responses to. HRV could provide a real time indicator of workload and provide visibility of staff effort to managers. It also remains essential to keep distinct what a physiological measure detects, and what is being inferred from the data about MWL.

How individuals differ in their physiological data could help explain why individuals differ in their subjective workload for a similar task. This has implications for MWL theory and subjective workload measures, beyond this rail specific study.

Chapter 6: Study 3 – Staff attitudes and perspectives on wearable measures

6.1 Chapter overview

This chapter presents the results from Study 3 the perspectives of staff to the use of wearable physiological measures. The interview study explored signalling staff perspectives on the potential use of wearable measures in the workplace in future. The study method combined semi-structured interviews and surveys with rating scales. Analysis considered to what extent personal attitude to change could predict technology acceptance. The study found wearable devices suit use in the live operational environment, with the wrist strap rated the most suitable due to low distraction and perceived ease of use. In terms of data use, themes included perceived usefulness, anonymity, and trust.

6.2 Introduction

The railway industry seeks new ways to measure and predict the mental workload (MWL) of signallers (Hack 2021, Zeilstra 2021) with the Rail Accident Investigation Branch (RAIB) recommending improved measures of signallers' workload (RAIB, 2020).

6.2.1 Staff Mental Workload

MWL is a construct that encapsulates task demand, individuals' experiences of MWL and task performance (Sharples, 2019). Staff MWL data can inform future design decisions regarding new assistive and automated technologies. This includes visibility of the impact on human performance of any technology introduced intended to increase operational performance and capacity. The aim is to monitor staff MWL and response to change, so that staff and operations remain within the boundaries of acceptable performance (see Rasmussen's (1997) Dynamic Safety Model).

6.2.2 Workload measures to suit live operations

Currently in the railway industry there is no single, widely used workload assessment method (RSSB 2005a). Measurement of MWL in live railway operational environments is a challenge, relying on short periods of an observer counting actions, or self-assessed ratings using the Integrated Workload Scale (IWS) (Pickup et al., 2005b), reported verbally to an observer. Measuring in the live environment is preferable, as the complex task of the signaller is difficult to realistically simulate, particularly for real-time communications (Sharples et al., 2011). Figure 6-1 shows a modern VDU workstation that includes automatic route setting, which is just one example of increasing task automation in this context (Sharples et al., 2011). This adds to the challenge of determining appropriate MWL. Reduced physical workload, due to the new technologies, presents two issues for MWL assessment:

1. The assumption by managers that when staff look like they are doing less work their MWL must be lower and their cognitive task is easy.
2. There are fewer observable cues that could infer an individuals' MWL.

In response to these issues, a simulation study was conducted to determine how wearable physiological data could supplement existing measures by providing visibility of workload in real-time to staff and their managers. This study found that these measures could provide continuous data of a signalling task with minimal distraction.

6.2.3 Importance of staff perspectives

The current study addresses the perspectives of staff to the use of these potential measures by interviewing staff about both the devices and the use of their physiological data. The importance of the perspectives of staff to the use of measures, and their data, became apparent during the thesis. As wearable measures become more accessible, the perspectives of staff to their use emerges as a research gap. This interview study provided a way to examine what contributed to staff perspectives.

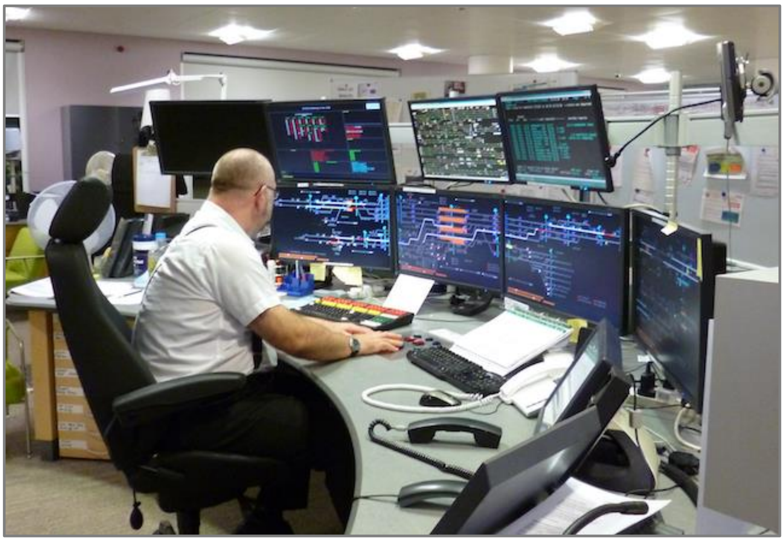


Figure 6-1 Signaller at VDU workstation, December 2018 (Source: Author)

6.2.4 Background

New physiological wearable measures offer an opportunity to track MWL in the live operational environment. Physiological state is balanced by the sympathetic and parasympathetic nervous systems in the human body. The sympathetic nervous system increases heart rate, alertness and sweating to prepare for action, whilst the parasympathetic nervous system decreases heart rate and increases gut secretions to rest the body (Sherwood 2013). Changes in mental state can be inferred from changing physiological state, with measures offering a "window into the mind" (Hugdahl 1995). Research has shown different physiological measures are sensitive to different aspects of MWL (Matthews et al., (2014), Charles and Nixon (2019)). Wearable physiological measures have been used in sports and health care, such as with athletes for mental acuity and stress (Seshadri, Li, Voos, Rowbottom, Alfes, Zorman and Drummond, 2019), rugby players (West, Williams, Kemp, Cross and Stokes 2019), remote medical health monitoring (Soon, Svavarsdottir, Downey and Jayne 2020) and personal fitness. Wearable devices are developing rapidly, for example the global fitness tracker market is projected to grow from £26 billion in 2020 to £83 billion[9] by 2028 (Fortune Business Insights 2021). Use of physiological measures in railway research has shown potential to assess driver and signaller MWL (Song et al., 2014, Broekhoven 2016, Crowley and Balfe 2018) but has, to date, been limited.

Electrodermal Activity (EDA) implies activation of the sympathetic nervous system and Heart Rate Variability (HRV) implies both sympathetic and parasympathetic nervous system activation (Tortora and Derrickson 2007). Alertness and attention can be inferred by EDA (Cacioppo 2017, Boucsein 2012) from a wrist strap (see

Figure 6-2 Wrist strap, Chest strap and App

). Increased workload can be inferred from low HRV (Lehrer et al., 2010) from a chest strap (see

Figure 6-2 Wrist strap, Chest strap and App

).

Figure 6-2 Wrist strap, Chest strap and App

Images: Wrist strap (Source: Author), Chest strap (Zephyr Bioharness, www.zephyranywhere.com) and App (Pickup et al., 2005b)

Previous research in rail found EDA inferred train driver alertness (Song et al., 2014) and anticipation (Crowley and Balfe 2018), and distinguished signallers'

[9] Based on an exchange rate 1 USD = 0.727 GBP, as checked on xe.com 23.10.2021

varying reactions to communications (Broekhoven 2016). In train drivers HRV decreased before and after stops, and at tunnels (Song et al., 2014). As a comparison to wearable measures, IWS is a self-assessed workload measure for signallers (Pickup et al., 2005b) (see

Figure 6-2 Wrist strap, Chest strap and App

). An app version of the IWS was developed by the Mixed Reality Laboratory, University of Nottingham for a simulation study of the signalling task (see Figure 5-4).

Despite the growth in wearable technologies, little research has been conducted on the acceptance of wearables (Gribel et al., 2016). The Technology Acceptance Model (TAM) is a widely applied information systems theory that explains technology acceptance and predicted usage. The Technology Adoption Cycle is a sociological model that explains why individuals vary in how swiftly they uptake new technology. TAM (Davis 1989) and its extension TAM2 (Venkatesh and Davis 2000) provide a scaled questionnaire to assess current attitudes, affording a valid and reliable way to predict future use of new technology. Whilst the original research focused on the use of software, it includes factors applicable to wearables. Figure 6-3 shows the factors of TAM2. Factors most applicable to wearables, as assessed pre-experience to predict Intention to Use are: Perceived Ease-of-Use, and Perceived Usefulness including Experience, Subjective Norm, Image and Job Relevance.

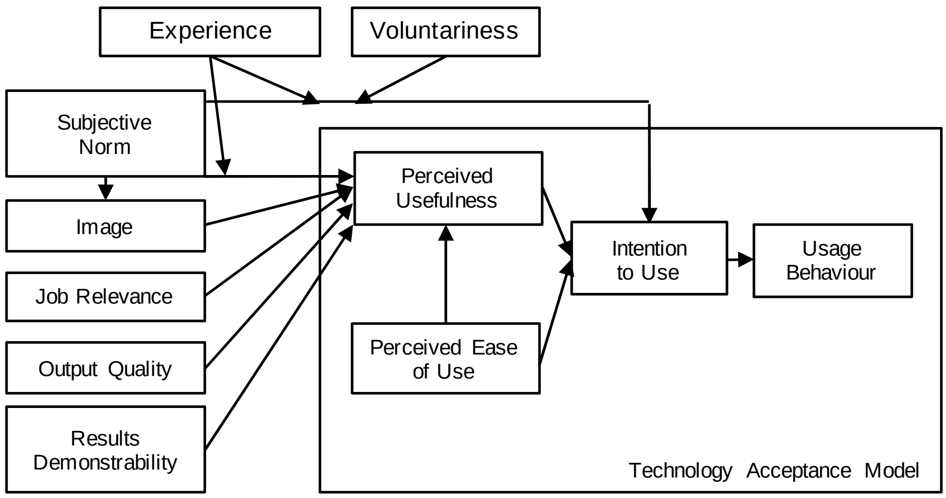


Figure 6-3 Technology Acceptance Model 2 (TAM2), Venkatesh and Davis (2000)

There are broader cultural reasons for acceptance or resistance of any new technology, captured in part by Subjective Norm in the TAM model. Limited research around resistance to technology has been conducted in the railway transportation domain (Rose and Bearman, 2013). A study of train drivers found negative opinions of new technology were be based on what individuals had heard from peers (Rose and Bearman, 2013), particularly when individuals had not experienced the technology for themselves. The study also found that whilst some staff saw the positives of sharing information with their manager to improve driving skills through training, other felt this was too 'big brother' and feared information would be used against them (Rose and Bearman, 2013). The latter points to more systemic issues of trust (Fox A, 1974). Staff need assurance as to how technologies will be used, and trust management that is how the technology is used. There are also concerns around the negative personal impact such as reducing the need for worker's skills without significant benefits (Rose and Bearman, 2013). These broader factors are potential barriers to acceptance that need to be managed. This includes ensuring staff are involved, that the intended use of the technology is communicated clearly, and that concerns are acknowledged. Resistance can reflect legitimate concerns such as safety implications (Naweed and Rose, 2018). Identifying which factors apply to the future use of wearables will assist in progressing their development, whether these be cultural, personal impact or safety concerns.

In addition to TAM2 that considers the functionality and cultural fit of the technology to the task, other theories consider the potential impact of individuals' attitudes to new technology. The Technology Adoption Lifecycle, a sociological model, describes the adoption or acceptance of a new product or innovation, and explains the varying speed of uptake across a population. The bell curve (Figure 6-4) is derived from the Diffusion of Innovations theory (Rogers 1962). According to Rogers' theory, innovators are the first to seek out new technology. Early adopters seek kindred spirits across industries. The early majority are pragmatists, they communicate more within their own industry and prefer being able to compare products. The late majority are conservative, they prefer to stick with what they know, only engaging with mature products. Laggards resist new technology. It is predicted that, in the rail industry, there may be resistance to new technology as staff are familiar with a slow uptake of new technology. The Individual Innovative scale (Hurt et al., 1977) identifies individuals' position on this bell curve.

In addition to TAM2, the following other factors should be considered when assessing attitudes and attitudes to wearables measures:

- Comfort (Urquhart and Craigon, 2021; Gribel et al., 2016; Wolf et al., 2018)

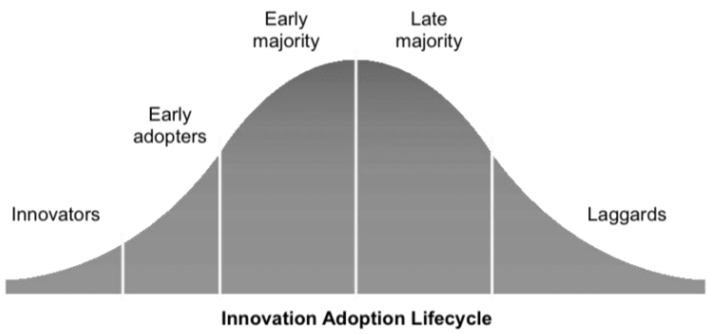


Figure 6-4 Technology Adoption Lifecycle (Moore 2002)

- Distraction (Jacobs et al., 2019, Gribel et al., 2016, Parasuraman and Colby, 2015, Urquhart and Craigon, 2021)
- Trust (Gribel et al., 2016, Jacobs et al., 2019, Parasuraman et al., 2015),
- Anonymity (Jacobs et al., 2019, and Urquhart and Craigon, 2021)
- Concerns over being tracked (Wolf et al., 2018)
- Sharing data with third party access (Gribel et al., 2016)

These factors were drawn from the Technology Readiness Index (Parasuraman and Colby 2015), the Moral IT deck (Urquhart and Craigon 2021), TAM2 adapted to wearables (Wolf, Menzel, Advisory, and Rennhak 2018, Gribel et al., 2016), Acceptance Factors of Wearable Computing (Gribrel et al., 2016), and staff attitudes to using wearables (Jacobs, Hettinger, Huang, Jeffries, Lesch, Simmons, Verma and Willetts 2019).

The current study determines whether wearable physiological measures could provide a timely addition to the toolkit of workload measures, by assessing current staff perspectives towards the use of wearable measures. Three measures are compared: a wrist strap to detect Electrodermal Activity (EDA); a chest strap to detect Heart Rate Variability (HRV); and the IWS as an app, to replace the previous method of verbal self-assessed ratings. These three measures were applied in a simulated signalling task. The inclusion of both EDA and HRV ensures both sympathetic and parasympathetic activity are accounted for. To assess railway staff perspectives on these new measures, this study draws on the interdisciplinary theories of the Technology Acceptance Model (TAM) and the Technology Adoption Cycle to predict likely usage of wearable physiological measures to assess staff workload, this study had the following research questions:

- Which measures suit live operations based on perceived comfort levels, perceived distraction from the task, and perceived relevance of data?

- What factors contribute to the perspective and attitudes of staff to the potential use of wearables at work?
- Can individual innovativeness or experience explain differences in attitude?

6.3 Method

6.3.1 Study design

The study applied a pragmatic approach (Robson and McCartan, 2015) by seeking stakeholders' perspectives and experience. The value of the findings was in identifying what works for industry, recognising the reality of the rail industry's operational setting is complex (Reichardt and Rallis, 1994). This study used semi-structured to explore how physiological wearable measures could be used in the rail industry in future. The use of semi-structured interviews allowed the stakeholders to share their experiences and opinions in their own word (Coveney, 2014). The method included elements of both an inductive and deductive approach. The semi-structured interview prompts supported a top-down deductive approach (Robson and McCartan, 2015) with stakeholder providing answers and examples on existing topics. The bottom-up inductive approach developed initially as the interviews progressed and additional topics emerged beyond the original prompts (Braun and Clarke 2012). The coding and analysis stages used a combination of both inductive and deductive approach to establish themes that reflected both the original deductive prompts and emerged during the interview process and coding. The themes that emerge from the interview data go beyond what can be observed (Glaus et al., 1996). The final themes and sub-themes inclusion reflect those relevant to the rail industry.

6.3.2 Participants

All eighteen participants recruited in this study were signallers and shift managers working at East Midlands Control Centre (EMCC) and included the local Union representative. Participation was voluntary, and no incentive was offered. Staff at the centre knew the researcher from the domain familiarisation activity undertaken during an internship with Network Rail (see 1.6 , and whilst arranging the live trials that were subsequently cancelled due to COVID-19 related restrictions in March 2020. These events assisting in gaining industry support for the interviews, both for participants and for the staff required to cover operations to enable the interviews to go ahead. The representative railway Unions were informed of the research ahead of interviews. Participants were recruited using a snow-ball sampling method, via a poster and managers.

The average age of participants was 46.9 years (SD = 12.9). Their average total signalling experience was 14.2 years (SD = 11.1), and experience of modern signalling VDU workstations was an average of 7.1 years (SD = 6.5). Higher Grade signallers were those at Grade 7, 8 and Shift Signaller Managers (who are qualified to cover workstations when required). Lower grade signallers were those graded 4 – 6. Participants represented 22% of the staff at one centre based in the Midlands. Participant numbers were randomly assigned, and do not reflect the order the interviews were completed in.

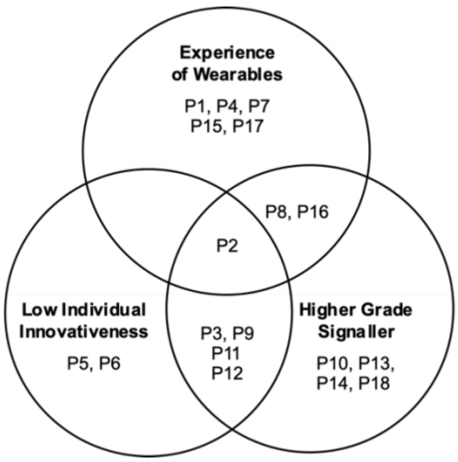


Figure 6-5 Participant experience of wearables, signalling and individual innovativeness

The Venn diagram in Figure 6-5 shows the distribution of participants based on whether they had experience of wearables for fitness, their Individual Innovativeness, and total experience of signalling. This demonstrates the participants covered a range of experiences from across all three of these demographics, enabling interviews to represent a range of perspectives and attitudes. The eight participants with experience of wearables for fitness reported using a wrist device, smart watch, or mobile phone app, and one who previously used a chest strap for fitness. Only six participants had experience of workload assessment techniques, spread across the three circles of the Venn diagram in Figure 6-5. All participants who had experience of reporting their workload using IWS were in the High Signalling Experience category.

6.3.3 Procedure

The questionnaire and interview questions (see Appendix D) were adapted from TAM and other existing methods and models (Venkatesh at al 2000, Hurt et al., 1977, Jacobs et al., 2019, Gribel et al., 2016, Urquhart and Craigon, 2021, Parasuraman et al., 2015, and Wolf et al., 2018). Unlike the technologies originally researched using TAM, the physiological measures investigated here are not designed to improve productivity directly (Jacobs et al., 2019).

In advance of the interviews, participants received a consent form, a short Pre-Interview Questionnaire to be completed online using Jisc Online Surveys, and a two-page pamphlet 'Introduction to Wearables and App'. This introduction contained pictures of all three devices (see

Figure 6-2 Wrist strap, Chest strap and App

), with a short description of each, to ensure all participants had the same briefing about the devices prior to being interviewed (see Appendix E). The Pre-Interview questionnaire collected demographic information, including

level of experience of signalling, workload assessments, and experience of wearables (see Appendix F). The Individual Innovativeness questionnaire (Hurt et al., 1977) was included to determined individuals' preference for new technology.
Use cases were referred to during interviews to present three alternative hypothetical future use cases that current staff could relate to (see Table 6-1). These were used as prompts to explore the reasons behind participants' answers on purpose. The case studies provided details of potential uses, such as examples of different levels of data sharing, and different purposes.

The Use Cases were developed by drawing on findings from the industry interviews study, the simulation study, the scoping review of physiological data, the literature on user acceptance of new technology and privacy, and in consultation with railway industry experts. They provided a range of:

- Perceived Usefulness (Venkatesh and Davis, 2000) - potential uses, based on industry interview of when staff are assessed currently (e.g. as a trainee, or by manager) and in consultation with industry. They include a workload assessment use, and two types of change: from trainee to being qualified signaller; and the impact of new technology.
- Job Relevance (Venkatesh and Davis, 2000) – most relevant aspects of workload to infer, based on the scoping review of physiological measures and simulation study findings.

Table 6-1 Use Cases used in interviews

	A **Understand Signaller Workload**	**B** **Learning Aid**	**C** **Assess impact of new technology or procedure**
Perceived Usefulness	Detect peaks & troughs in workload & effort	Track progress, self-learning, assess training	Assess effectiveness of change
Job Relevance	Anticipation, alertness, stress, time pressure, brief peaks	Alertness, confidence, unexpected event, stress, effort	Stress, effort, unexpected system responses
Anonymity & Trust	Anonymised	Trainee data shared with trainer & device supplier	Labelled with initials, shared with manager & investigator
Time required	Data collected during 1-2 shifts	Data collected across training period	Data collected across period of change

- Anonymity and Trust – level of anonymity (Jacobs et al., 2019; Urquhart and Craigon, 2021) and third party access (Gribel et al., 2016) from the scoping review of privacy in personal data use.
- Duration – how long would it be feasible to collect data for, based on consultation with industry.

The semi-structured interviews were conducted over the telephone and audio recorded. The interview prompts are presented in Appendix G. The average length of interview was 60 minutes (SD = 14 minutes), with a total of 18 hours recorded. Interviews could not take place in person due to COVID-19 restrictions stopping visitor access to operational sites. All interviews took place during work hours, with cover arranged at the workstation to allow staff to participate. The study was approved by University of Nottingham's Faculty of Engineering Ethics Committee, as shown in Appendix D.

The short Post-Interview questionnaire asked participants to rate, on a scale, their level of agreement with eleven statements on topics covered in interview (adapted from Venkatesh at al 2000, Jacobs et al., 2019, Gribel et al., 2016, Parasuraman et al., 2015, and Wolf et al., 2018). The questionnaires included scale questions from 1 strongly disagree to 7 strongly agree (see Appendix H). This compared responses between the chest, wrist, and app devices.

6.3.4 Data Analysis

An Automatic Transcription Service was used to initially transcribe interviews. Familiarisation was achieved during checking and manual correcting of all interviews by the first author of the study.

Thematic coding was used in an iterative process to discover the factors underlying staff perspectives and attitudes (Robson and McCartan 2015, Braun and Clarke, 2012, and Saldaña 2016). Half the interviews were coded on paper to produce the framework v1. A coding review exercise was performed at this stage, in which a second analyst sorted the child codes into thematic groups, without reference to framework v1. All child codes were retained, with 38 matching their eleven existing parent codes. The results of this check were examined by the two analysts, which resulted in some parent codes being merged and a new parent code ('Data Quality Uncertain'). This stage is presented in Figure 6-6, showing parent and child codes.

After the check by two analysts, all child codes were merged into their parent codes. The resulting revised framework v2 comprised of 10 codes. Coding was digitised using NVivo 12 (QSR 2019) and all interviews were coded. Two codes were merged so the final framework (presented in Table 6-2) contained 9 codes across three stages: Justification, Data Collection and Consequence.

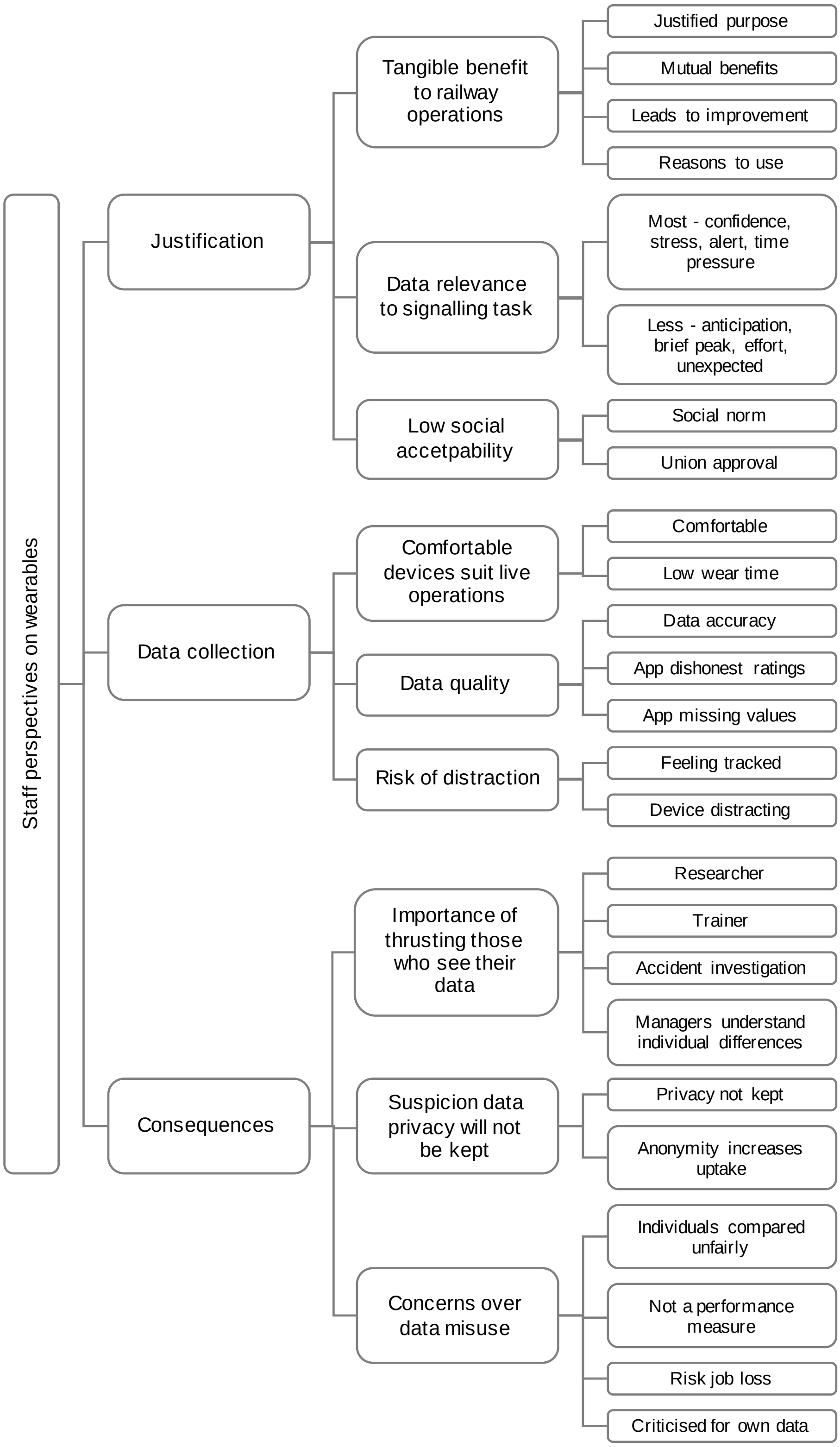


Figure 6-6 Coding tree for staff perspectives on wearables

Independent of the coding exercise, the researcher devised a method to categorise the results from the 32 interview questions. The categories indicated the extent to which participants supported the use of measures. A traffic light coding was applied to each answer provided. Red indicated that participants' answers explicitly contained responses that were against the use of the measures. Green indicated participant responses explicitly endorsed the use of the measures. Amber indicated responses that were mixed, such as containing caveats to their use, or where they did not wish to comment either for or against the measures. This analysis was specifically used to determine whether a pattern existed between experience, or individual innovativeness and individual perspectives and attitudes to the new measures, and robustness in the analysis.

6.4 Results

The iterative thematic coding produced 9 codes, across 3 stages, and represented a range of perspectives and attitudes from endorsement, uncertainty, to opposition to the use of wearable measures (see Table 6-2).

Table 6-2 Coding framework for perspectives on wearables in the workplace

	Stages →		
Attitudes	**1 Justification**	**2 Data Collection**	**3 Consequences**
Positive ↑	Tangible benefit to railway operations	Comfortable devices suit live operations	Importance of trusting those who see their data
	Data relevance to signalling task	Data quality uncertain	Suspicion data privacy will not be kept
Negative ↓	Low social acceptability	Risk of distraction	Concerns over data misuse

The themes identified in the study are presented in the sunburst chart in Figure 6-7, and described in the following sections, presented with results combined from interviews and questionnaires.

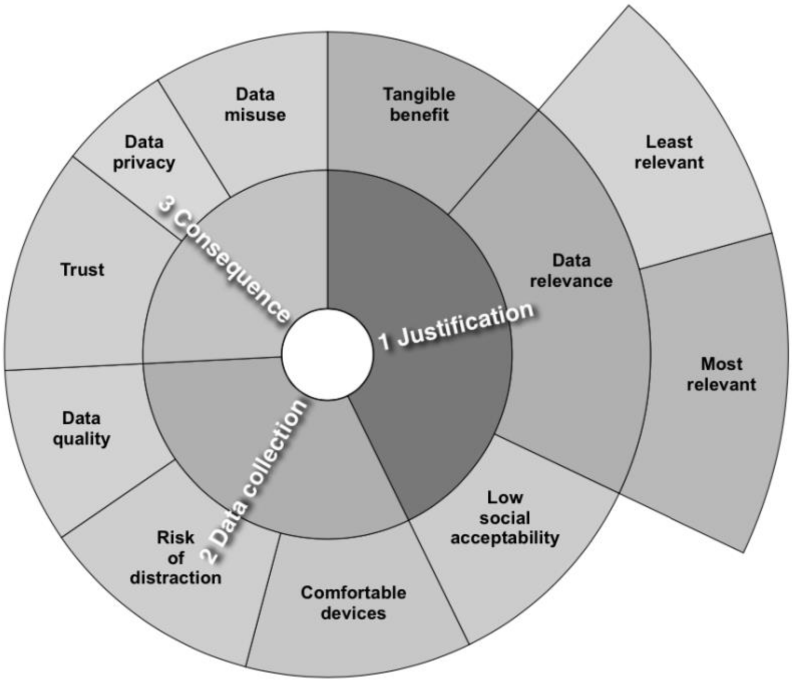


Figure 6-7 Sunburst of framework, with distribution of number of items coded.

6.4.1 Stage 1: Justification

This first category is comprised of codes on the benefits and relevance of measures and their social acceptability. All codes relate to how well wearable measures currently fit into the railway industry.

Tangible benefit to railway operations

Participants overall held positive views that measuring signaller workload would be useful in one, or more, use cases to: demonstrate to others how hard they work; understand their own data as trainees; or demonstrate to others the impact of change. Questionnaire responses matched these findings, with participants agreeing that measuring workload is important in the railway industry (Mean = 5.61, SD = 1.88), on a scale 1, strongly disagree to 7, strongly agree. Wearable measures were seen as having potential, provided they had clear operational benefits [P13, P14, P15, P15] that led to improvement in operations [P7], effectiveness [P3] or safety [P7, P18]. Trusting the use of the data (covered in a later section of the results) would be improved if the company was clear on what precisely they hope to achieve with wearables [P4, P5]. One potential benefit was improving the visibility of effort staff put in, exemplified by the following quotes from interviewees:

“The job looks easy to somebody sitting outside … When it looks like you're doing nothing because you’re on top of (the) task.” [P14]

“I think it is very useful in certain senses. I think people in different areas of rail operations, we don't see what they do, and what their stress levels are, and they don't see what the signallers’ side is.” [P8]

Participants felt wearables would provide objective data, with fewer false readings [P3, P5], and more overall readings than the app, which may not be completed during an incident when workload was high [P16]. Overall, participants were receptive to a live trial of physiological measures, to provide evidence of the data that wearable measures could offer.

“I think if we're looking for ways to improve efficiency or improve safety than then it's a no-brainer.” [P7]

Participants suggested measures could show the variance in workload across a shift, across days, and workstations, and the varying pressure [P18] that may not be obvious to others [P9]. Measures may also show when staff are stressed [P6] or fatigued and require a break [P1]. Participants thought it would be interesting to see how trainees react to certain situations [P12, P17], their stress levels [P7, P8, P12, P18], concentration levels [P8], attention [P18], alertness [P10, P16], and confidence [P5, P7, P16], compared to experienced signallers [P8, P18]. Measures could assist in training and assessment [P6, P18]. Data could also inform a debrief activity [P15], to reflect on situations [P15, P16], and for the trainee to become more conscious of what they are doing [P2], when they may not be thinking about how they are performing [P18]. Potentially both trainer and trainee could benefit from wearing a device [P16] and comparing outputs. Regarding the impact of change, participants commented that not all changes were better

[P7, P8, P15]. Data may provide a more accurate picture of the impact of technologies, whether they work as intended [P16], and to adapt the roll out [P5] if not. In addition to the use cases, participants commented that the impact of incidents on staff is not known. Measures may be able to assess the effect incidents have on staff, particularly incidents such as a fatality when a train hits someone[10] [P17].

Data relevance to signalling task

In the questionnaire participants agreed that measuring their workload is relevant to their job (Mean = 5.61, SD = 1.82). To determine whether the chest strap or wrist strap were relevant to signallers' workload, participants were presented with examples of what could be inferred from physiological data. Stress, confidence, and alertness were deemed most relevant to signallers. Anticipation and time pressure were deemed less relevant. Stress and alertness can be inferred from the wrist strap and confidence or uncertainty, from the chest strap.

Participants reported that stress can occur with failures, incidents, or emergencies [P2, P4, P8, P9, P15, P17, P18]. For example, points failure [P2], bridge strike [P3, P5, P9], children on the track[P3], or trains disappearing off the screens due to leaf fall[11][P6, P9]. Drivers only call when there is a problem, so hearing the phone ring can be stressful [P1, P2, P9] until the problem is known [P1, P5, P9, P14, P16]. Incidents can then take time to wind down from: *"All of a sudden you've got an emergency call to deal with and you're finding out whether or not a driver/ a train has actually run over someone."* [P1] Not all Unexpected events are stressful. Once the signaller knows what to do [P9], some events less stressful [P11, P17], such as a tree [P17] or cow

[10] Most fatalities are suicides. Anyone affected by suicide can gain support by calling Samaritans, UK, 116 123 or NSPI, USA, 1-800 273 8255; SRPC, Beijing, China, 0800 8101117; or AASRA, Mumbai, India 91 22 275 46669. Additional international helpline numbers are available via Therapy Route (2020)

[11] Leaf fall, compressed by trains can, in track circuit areas mean the metal wheels lose contact with the track.

[P13] on the line. Brief peaks in MWL in these circumstances would not affect signallers for long [P6, P8, P11, P12, P14].

Participants reported feeling confident when everything runs smoothly and on time [P1, P2, P5, P8, P9, P12, P15]. Trainee confidence builds with experience over time. Once they successfully deal with something the first time, they know they could a second time [P4, P6]. As their confidence increases, their stress decreases [P2, P5, P13], and they deal better with time pressure [P13] and when unexpected events occur [P8]. *"Confidence. I believe it's a big thing in this job. And knowledge is confidence. You know if you've got the knowledge, you ARE more confident. And that gives you a much better set of skills to work the workstation."* [P15]. Confidence also links to alertness. Trainees start with a heightened level of alertness, trying not to miss anything [P7], whereas, experienced signallers can sit back, observe, whilst being prepared for problems. For example, a participant with high experience gave an analogy to compare a trainee with a seasoned signaller: *"Because you're new in the role you're there like a meerkat, because you're constantly looking, you're always then anticipating what's going to happen next? [Compared to] an elephant or rhinoceros or something like that that's possibly not as threatened by predators, something that's more laidback. So, the alertness level is there because they ARE looking for the dangers around them, BUT they're not up on their hind legs scouring"* [P16]. Other participants suggested that trainees may have heightened levels of alertness, describing them as having 'eyes on stalks'[12] with alertness 'through the roof'[13] [P5], with those new to the live workstation in a constant state of fear for the first few days [P7, P14]. These findings suggest confidence builds with experience, and with this, alertness is reduced from a very high level to a regular level. The trainee gradually gains sufficient experience and confidence to turn from the 'meerkat' to the 'elephant'.

[12] An English phrase meaning their eyes are wide open with surprise.
[13] An English phrase meaning to rise to a very high level.

Time pressure and anticipation were deemed less relevant to signallers. Regarding time pressure, signallers acknowledged their job is ‘all about time’ [P12]. One participant said, as an example, *“I must look at the clock more than anything else in this place.”* [P18]. Not all participants viewed this as time pressure [P6, P13], as staff manage their own workload [P1, P15]. There may be, however, multiple people waiting for a line block [P2, P4, P5, P6, P14], or only seconds to enter a route for a busy junction [P9], or minutes to give up a possession and allow trains to run again [P12]. This seems to suggest that, whilst time pressure is a daily external task demand, this does not automatically lead to a negative experience of pressure. Ensuring the timeliness of train services is a signaller responsibility. It could be this does not automatically translate into time pressure for signallers as they have the control to delay their tasks, if that is what is required to complete their tasks safely. Signallers can inform staff on site they must wait to start work or they can keep a train held at a red signal.

Anticipation was viewed as bad for a signaller [P11, P13] when the workload cannot be anticipated such as the phone ringing. Signallers plan what they can, e.g. routing trains and regulating the order trains proceed. Their responses suggest they do not view these as anticipation. They suggested that they instead indicated anticipation is a risk if it is based on poor assumptions of what normally happens [P13] or dwelling on what might happen [P12]. Participants advocated a level of preparedness instead, to notice issues when they occur, and deal with them promptly. *“You can NEVER expect to have a problem. It's just knowing how to deal with that problem at that time.” [P16]*. Preparedness included being aware of weather conditions [P14] and the likely workload of a workstation [P5, P12]. Noticing issues, included regulating (deciding the order of trains) [P2, P5, P8, P10, P11, P16, P17] and noticing, on CCTV, those pedestrians or vehicles that may not comply with level crossing lights and barriers [P7].

Low social acceptability

In interviews, despite signallers themselves seeming prepared to use at least one device, they anticipated a mixed response from colleagues. When asked 'What is the likely reaction of colleagues to the use of these new measures?' only 17% of participants responded that the reaction of colleagues would be positive. The remainder anticipated some level of resistance to the new measures, including the wrist strap, chest strap and app. One participant's comment sums up the expected range of responses: *"I think we've got those that don't like change in any way, shape or form, they won't like it. You'll get those who'll think 'Oww, you know, this is something new I'd like to try this'. Then I think you'll get the middle of the road 'So, well, if it does me any good then I'll give it a go'".* [P18] Resistance seemed due to concern over what the information could be used for, and fear, such as fear they might lose their job. *"This is a backhanded way of the company trying to monitor us ... The union will be all over that."* [P11]. This fear, and associated resistance, present a barrier to the future use of these measures.

In response to the precautionary principle (Urquhart and Craigon, 2021) 'Just because we could use these measures, should we?', nine responded endorsing the measures, 9 were uncertain, and none were against the measures. Predicted use and intension to use resulted in the same ranking of measures. The wrist strap had the highest predicted use, on a seven-point scale with 7 high (wrist strap Mean = 6.06, SD = 1.00, app (Mean = 5.28, SD = 1.56, chest strap (Mean = 5.11, SD = 1.68) (see Figure 6-8).

In the questionnaire signallers, on average, slightly disagreed with the statement: 'I would not recommend the devices to my colleagues' (Chest Mean = 3.33, SD 1.37, Wrist Mean = 3.00, SD = 1.53, App 3.22, SD 1.40). This shows that, despite signallers expecting a mixed response to measures, they would somewhat be prepared to recommend devices.

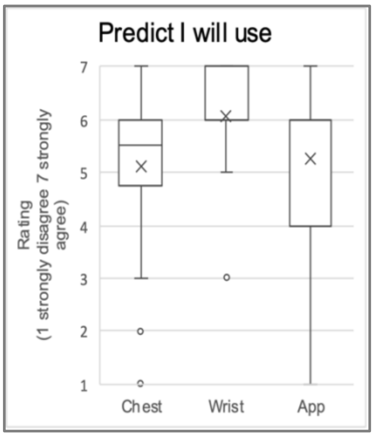


Figure 6-8 Questionnaire responses to prediction of use of wearables.

Box shows interquartile range, x is the mean, line the median, whisker is 1.5 times interquartile range, and dots are outliers.

Regarding the TAM2 component 'image', signallers somewhat disagreed with the statement in the questionnaire: 'The devices could be a status symbol in my organisation' (Mean = 2.94, SD = 1.43). Wearables were not viewed as improving image, with one signaller commenting that there could be *"a little bit of banter"* from colleagues for wearing one [P3].

6.4.2 Stage 2: Data collection

This second category is comprised of codes on comfort, data quality and distraction. All relate to devices and the data collection stage.

Comfortable devices suit live operations

Comfort was associated with increased experience and reduced awareness whilst wearing. *"I wear one (an apple watch) all the time, don't I so I don't really think about it to be honest. I'd be happy to wear one all shift, for research* [P8]. Signallers were familiar with wearing something on their wrist [P16], like a watch [P7] that they wear all the time [P17], or Fitbit [P1]. Once the watch is on, they don't really know it's there [P1], it becomes 'part and parcel' [P16] and they forget about it [P12].

"It's like anything that's new. You've got to take time for it to become part of the norm." [P16]

Signallers commented the chest strap was a bit strange [P3, P5], would require "*more messing about putting it on*" [P3], and may be an irritant [P8] or uncomfortable [P12], but it was difficult to predict comfort. All except one signaller had no experience of chest straps. *“I don't know, (laughs) but obviously without trying without trying the chest one it's hard to say”* [P4]. The signaller who was familiar with the chest strap was less aware of it: *“I've used them (chest strap) before for like heart rates you know and fitness type stuff, and once you get going, you kind of forget about them”* [P1]. Whilst signallers may initially be conscious that they were wearing them, this would be likely to reduce once they were used to them [P13].

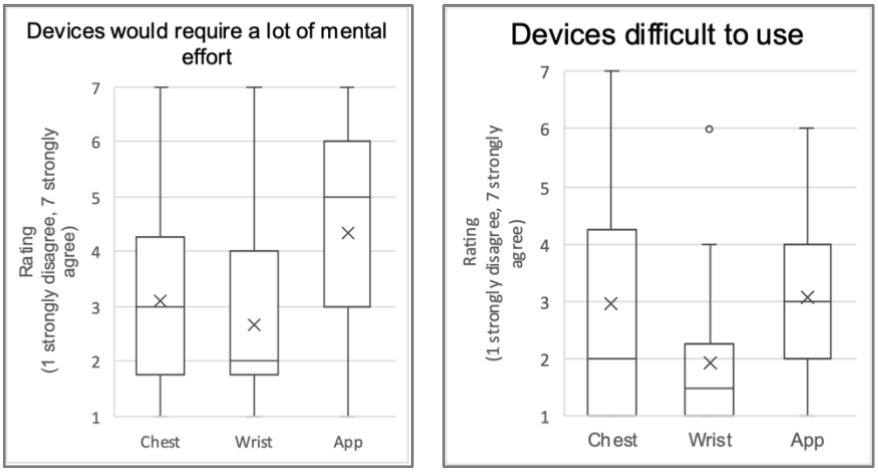


Figure 6-9 Questionnaire responses to devices being difficult to use or require mental effort

Box shows interquartile range, x is the mean, line the median, whisker is 1.5 times interquartile range, and dots are outliers.

Figure 6-9 illustrates the results for the questions, ‘Wearing devices wouldn't require a lot of my mental effort’, and ‘I would find the devices difficult to use’. The questionnaire data indicated a trend to the wrist or chest strap requiring less mental effort (Wrist Mean = 2.67, SD = 1.61, Chest Mean = 3.11, SD = 1.88) than the app (Mean = 4.33, SD = 1.75). This is consistent with the app requiring signallers to enter ratings compared to wearables recording data passively. All devices, on average, were rated as not difficult to use, with the wrist strap being the lowest difficulty (Mean = 1.94, SD = 1.35). The three signallers with rating scale experience rated the app as difficult to use (ratings of 6, 4 and 6). The wrist strap overall reportedly suits use in the live environment. The chest strap may suit use after a period of familiarisation. The app was rated the least suitable, as it requires the most mental effort.

Data quality uncertain

The data quality code includes concerns by participants that firstly app data is reduced by staff missed ratings [P1, P2, P3, P10, P11, P12, P15], particularly when busy and potentially for a while after an incident [P16]. *"If you are going to pursue the app, you probably are going to have to accept that some of the data is going to be quite sporadic"* [P11]. Secondly, that data from the app may be skewed [P16], depending how honest people are [P12]. People won't report when they are too busy, they will under-report [P5, P11, P12, P15, P16], or overreport pretending they are busy [P5, P11, P16], enter a value without much thought [P6], or different individuals will report different values for the same situation [P11]. In the eyes of the signallers all these points make the app data less reliable in the eyes of signallers. Finally, participants queried how accurate physiological data would be for workload. For example, how to differentiate individuals who appear stressed, and others who do not [P11], or when self-assessed workload does not match physiological data [P16]. Participants understood that stress increases heart rate but queried whether it was clear what was good or bad levels in terms of workload [P5].

Risk of distraction

Participants thought the app the most distracting and the wrist strap the least distracting. "*I don't think the wrist one would be an issue at all*" [P4]. "Sometimes there's three phones ringing at the same time, and you've got to decide which one you want to answer. Never mind just interact with an app" [P9].

Participants thought the app would distract staff from their *job [P7, P8, P13, P14*], in particular the sound [P4, P9] which may even distract staff at another workstation [P1, P5]. Mobile phones are not allowed on the floor due to the risk of distraction [P7, P13]. The app would be especially unsuitable for trainees [P4, P10, P11, P13, 14, P18] who may be at risk of overload [P14] or may interact with the app and miss something [P9]. *“On a live workstation, we've got safety critical situation, we're asking them to possibly, just by doing this (answer the app) to distract them from what they're already doing”* [P14] In interviews participants reported the chest strap may be more intrusive than the wrist strap [P2, P8] and they would be more conscious of wearing it [P2, P18]. Two participants added they may also be distracted wondering what the data was showing [P2], particularly if there was an incident [P7]. Despite this, on average, participants were not concerned about being tracked, with low ratings for the statement ‘*I wouldn't use the devices because I would be concerned about being tracked*’: chest (Mean = 2.06, SD = 1.11); wrist (Mean = 2.06, SD = 1.11); and app (Mean = 2.00, SD = 0.97).

The questionnaire confirmed the most distracting was the app (Mean = 4.72, SD 1.56), then the chest strap (Mean = 2.94, SD = 1.73), and the least distracting device was the wrist strap (Mean = 2.50, SD = 1.38) (see Figure 6-10). In the questionnaire, those who agreed or strongly agreed that the app would be distracting [P5, P9, P13, P18] had no prior experience of wearables. One had previous experience of the IWS.

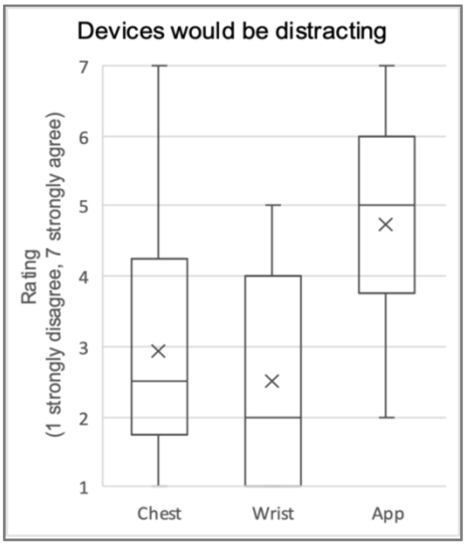


Figure 6-10 Questionnaire responses to devices would be distracting

Box shows interquartile range, x is the mean, line the median, whisker is 1.5 times interquartile range, and dots are outliers

6.4.3 Stage 3: Consequences

This third category of comprised of trust, data privacy and data misuse codes. All relate to what happens with the data collected from devices.

Importance of trusting those who see their data

Level of trust in sharing named data varied depending on who sees the data. Participants were more inclined to trust those external to the organisation: researchers who process the data before anonymisation, and device suppliers who need access. Participants were more cautious about sharing named data within their organisation. Over half the participants agreed trainers could see trainees' data. Only 11% of responses were positive regarding sharing named data with managers. One reason given was managers may misinterpret data due to lack of understanding [P1, P5]. *"If they ((managers)) don't fully understand what's going on, then I'd rather you ((a researcher)) come and tell me"* [P1].

Sharing data with accident investigators met with mixed attitudes. Some participants queried how data would be used [P13] and its relevance [P5], whilst others accepted it [P6, P12, P15, P18]:: *"Honestly really is the best policy … in regard to the last question about being shared with investigators, I think another reason I don't think it (would) be that much of an issue, I don't think you should be working if you're not in the right state of mind"* [P6]. Two dichotomies were identified regarding participants preference for anonymising data. Firstly, for those wishing to see their own data, it must be identifiable so could not be fully anonymised. Secondly, while increasing anonymity may increase acceptance, anonymisation is a challenge, such as for trainees when there are only a few at one time [P7], as date and time, and workstation data could be used to identify an individual.

Suspicion data privacy will not be kept

Participants were asked whether they had any concerns about data use. Seven raised concerns around data privacy. One aspect of this was word spreading [P1, P6], despite steps to anonymise data: *"You don't want to walk into work and then everyone saying 'Ohh we've seen your stress levels yesterday'"* [P6]. It could be particularly hard on trainees if they were talked about. Yet it is a challenge to anonymise the data, as rosters show who is on shift [P2] and, if only a small number of participants wear measures it would be obvious who they are [P14]. Both these confirm anonymisation would be difficult in practice.

Concern over data misuse

Fourteen participants raised concerns that data may be misused by managers. Their ultimate concern was the risk of losing their job [P5, P11, 17]: *"Oh the company will get rid of me then if they don't think I could do the job properly"* [P11]. They didn't think colleagues would want this added pressure [P11]. Participants were concerned managers would criticise staff about their physiological data [P2, P9, P7], as has happened with delay attribution [P2], or wrong routes [P1]. Participants did not want the data used for disciplinary matters [P16], to assess job performance [P11], or as a negative indicator of their capacity [P18]. Another concern was staff may be unfairly compared to each other [P5, P9, P11, P13, P14]. Staff heart rates differing [P5], or differing levels of stress [P5, P13], were not seen as relevant if the performance outcome was successful [P5]. It was deemed unfair to compare trainees in training, between someone working on their own, with someone who has a trainer with them [P13]. If, instead, the data was not used to incriminate staff in any way, and kept only for research, then staff should not have a problem [P17]. *"I'm all for it and I'm quite happy with the technology, as long as they are USED by the company in the correct way. That's vital that's vital"* [P14].

6.4.4 Individual Innovativeness and experience impact

In addition to the codes in the Framework, the demographics of level of individual innovativeness, and signalling experience, were analysed to determine if they predicted variation in attitudes. This found attitudes did not predict proportion of interview responses endorsing new measures. Results showed a slight trend to low signaller experience being more positive about measures (Mean = 69.0%, SD = 7.9) compared to a high signalling experience (Mean = 57.6% (SD = 21.7). A pattern that did emerge was that signalling experience and individual innovativeness were linked with experience of wearables for fitness. All low signalling experience, high innovativeness (n = 5) had experience of wearables, compared to all low signalling experience, low innovativeness (n = 2), who had no experience of wearables. High signalling experience had varying experience, (3 had experience, 8 had none).

6.5 Discussion

To predict the likely usage of wearable physiological measures to assess staff workload, this study assessed: which measures most suit live operations based on perceived comfort levels, perceived distraction from the task, and perceived relevance of data?; which factors contribute to staff perspectives on the potential use of wearables at work?; and, can individual innovativeness or experience explain differences in perspectives and attitudes?

6.5.1 Suitability for live operations

The device most suitable for live operations was the wrist strap, being the most comfortable, least distracting device, that staff would be unaware of wearing. Participants preference for the wrist strap could be explained by familiarity with wrist watches, compared to the novelty of wearing a chest strap. The app raised the greatest concerns over distraction, particularly for trainees. These findings support the previous theories and models that suggest comfort (Urquhart and Craigon, 2021, Wolf et al., 2018), and distraction (Jacobs et al., 2019, Gribel et al., 2016, Parasuraman et al., 2015 and Urquhart and Craigon, 2021) are valid factors in staff attitudes to wearables at work. Regarding the most relevant data, stress, confidence, and alertness were deemed most relevant to signallers. Anticipation and time pressure were deemed the least relevant. Whilst stress and alertness can be inferred from wrist strap data (Healey and Picard 2005) confidence, or uncertainty as a lack of confidence, is inferred from the chest strap data (Ramírez et al). If confidence is to be tracked, a chest strap could be considered, after a familiarisation period.

6.5.2 Factors contributing to staff perspectives on wearables

A key factor that emerged that contributed to staff perspectives and attitudes toward the use of wearables was whether they trusted the staff who would see the data, and what the negative consequences could be of sharing data, particularly if named data is shared. Trust, as a key part of gaining staff agreement, matches theories and models in the literature (Gribel et al., 2016, Jacobs et al., 2019, Parasuraman et al., 2015). No devices would be worn by staff if they were concerned managers would use the data to criticise them, blame them, or assess their job performance. The ultimate concern was misuse of data by managers leading to staff losing their job. The alternative to this was that a clear operational benefit of the data, leading to improvement in operations or safety, would encourage staff to trust the use of their data. Regarding whom specifically they would trust, participants' responses suggest they would trust researchers outside the organisation to interpret the results and pool data across participants to protect individual identities. The distinction between trusting those outside the organisation, compared to managers, may reflect the relative control over consequences that these parties have. Managers have the most influence on supporting or disciplining staff. External third parties, with less direct control, may be seen as less of a threat. This concern is removed if data is fully anonymised, matching existing theories (Jacobs et al., 2019, Urquhart and Craigon, 2021). One dichotomy here is, within a signalling centre, anonymisation would be difficult to achieve, as workstation and date would be sufficient information for staff to know, from the roster, who was on shift. Another dichotomy is individual staff could not see their own data if it was fully anonymised.

Three factors identified in theories that did not appear to contribute to staff perspectives and attitudes of wearables were devices being seen as a status symbol (Venkatesh et al., 2000), concern over being tracked (Wolf et al., 2018), or sharing data with third parties (Gribel et al., 2016). Participants did not view the measures as a status symbol, possibly as they do not directly support performance outcomes, and are not viewed as a privilege. The data from the wearables was viewed instead as a potential threat due to the risk of misuse. This did not show as a concern about being tracked. This may be due to staff in rail being used to high levels of existing monitoring (such as phone calls content, workstation inputs logged).

6.5.3 Individual Innovativeness and signalling experience

Individual innovativeness or signalling experience did not explain differences in perspectives or attitudes. Those familiar with wearables did not automatically endorse the use of wearable measures. Familiarity with wrist watches however, may explain a preference for a wrist strap over a chest strap. In addition, all those with high innovativeness, low signalling experience, had experience of wrist wearables and mobile apps for fitness compared to only 3 out of eleven participants with high signalling experience. These findings, taken together, suggest the proportion of individuals with experience of wearables will increase as more staff join. This is predicted to be higher if those individuals have high innovativeness. If staff join with experience of a chest strap, they may be more likely to wear a chest strap at work without the device being a distraction. So experience may influence attitudes (Gribel et al., 2016, Jacobs et al., 2019, Venkatesh et al., 2000), but in this study it did not predict attitudes.

6.5.4 Implications of findings for understanding signalling workload

This study has implications for understanding the individual experience of MWL by signallers. Firstly, individuals' experience of workload is not the number of trains on a workstation. The number of trains timetabled varied across workstations. Staff seem to prepare themselves, prior to a shift, for level of task demand. This would fit with the malleable MWL theory of Young and Stanton (2002). Instead, changes in stress and alertness, were due to either unexpected, or novel external events such as incident and associated individual internal factors such as level of personal experience and confidence. Novelty and uncertainty seem inextricably linked, with confidence from experience a counterbalance to this.

Trainees start stressed and with high alertness. Over weeks and months this subsides as their experience and confidence build, as they successfully respond to a range of unexpected events. This includes signallers building experience of making others wait (e.g. trains or staff), in order to manage their own workload and control time pressure. Physiological data could provide an indicator of the resulting changes in the underlying state, and the balance between sympathetic and parasympathetic nervous system. The dichotomy here is that participants acknowledge trainees may have heightened alertness, compared to experienced signallers, but felt comparing results would be unfair, and trainees' data could falsely represent the workload imposed by the task. Participants felt comparisons should only be on operational performance outcomes, not individual experiences of MWL. This caution may be linked to participants' concerns over the misuse, or misinterpretation, of physiological data. These concerns would need to be addressed before physiological measures are applied, to understand this variation in individual experience of MWL.

6.5.5 Generalisability in rail and other industries

How generalisable these results are to other staff roles and tasks in rail, and other industries, will vary. This study suggests the greatest changes in experienced MWL for signallers comes not from the number of trains but from how novel events are to them and the associated additional tasks required including phone calls, increased regulating and managing disruptions. In other countries, these findings could be generalisable to controllers who play a similar strategic role recovering from disruption (Dorrian et al., 2011). Similarly, the results are generalizable to other control roles in safety critical industries. The results may be less applicable to the more tactical role of signallers in parts of the world where their role is more implementing a plan. Where the strongest generalisability exists is where there are conceptual similarities in role to the signallers here, rather than the association being purely by domain.

The extent to which the results map to train drivers is unclear. Anticipation was not identified as relevant to signallers. This contrasts with previous research of train drivers which did identify anticipation as relevant (Crowley and Balfe, 2018). Where the physiological data may be generalisable to drivers is around inferring stress and alertness from EDA, and confidence from HRV data (by an absence of uncertainty). To select the most applicable wearables for each role in rail, and each industry, would need to determine which physiological responses were most relevant to understand the MWL of that specific staff role.

Each role, or each industry may prove to be at a different stage of securing appropriate Justification, Data Collection and Consequences.

6.6 Limitations

All participants scored above average on the Individual Innovativeness scale. The results of the study may be biased towards a more positive perspectives and attitudes on the use of new measures. Voluntary participation may have led individuals with lower innovativeness to not participate in a study about new technology for future use. Alternatively, this finding could indicate signallers working with the most modern signalling control systems, the VDU workstations, are naturally biased towards high innovativeness.

6.7 Conclusion

This study considered the perspectives and attitudes of staff to the potential use of wearable physiological measures to detect signaller MWL. The study suggests wearable devices suit use in the live operational environment, with the wrist strap rated the most suitable due to low distraction. Physiological data from the wrist strap could provide visibility of individual MWL of staff, in particular their stress and alertness, both relevant to signalling. Such data could build our understanding of individual workload, across a range from underload to overload, and from novelty to familiarity. It is essential however that staff trust those who will see their data, particularly if they share named data. This trust would need to be in place before staff accept wearable measures. One way this trust can be built is through clear operational benefits that lead to improvement in operations. A certain level of resistance to measures is to be expected due to this low current experience and the novelty of the devices. If the chest strap was needed, a period of familiarisation would be required. Whilst this study focuses on railway signallers, the findings have implications for other roles where human performance is key to the control and monitoring complex safety systems.

Chapter 7: Discussion and implications

7.1 Chapter overview

This chapter presents answers to the research questions and novel contributions of this research. The discussion draws together findings from the industry interviews, simulation study, and perspectives and attitude studies. The discussion includes how physiological measures can contribute to MWL assessment, staff perspectives and attitudes on their use, theoretical implications, and implications for industry.

7.2 Introduction

This research considered how temporal physiological data from wearable measures could monitor mental workload. The industry interviews identified MWL, including underload, as an industry risk with increasing automation technology. The research aim was to progress the measures for use in live operations, in real-time, to assess staff mental workload with minimal task interference. An overarching ethos of this research is to consider positive human performance and explore 'what does good look like', rather than focusing only on when errors are made. The research aimed to detect not only MWL underload or overload but also when MWL is between these, so is more sustainable and supports successful operational performance. The research scope considered how personal data could be used in rail to measure human cognitive performance to, in future, provide feedback through visibility to staff and their managers. This would inform decisions on how best to manage staff workload to support successful operational performance.

The temporal data from physiological measures highlights dynamic changes and provides a chronology of events to provide a timeline of a continuous task. This data can be used to visualise patterns that could indicate when staff are at risk of moving from successful performance into the higher risk areas of either underload or overload. The findings suggest that this risk of overload is higher when there is event novelty. The discussion centres around answering the three research questions:

1. How can temporal physiological data from wearable measures contribute to MWL assessment in rail industry live operations?
2. What are the theoretical implications of individual physiological data to changes in MWL in a workplace setting?
3. What are staff perspectives on wearables and use of their personal physiological data?

Whilst the results focus on railway signallers, the implications of the findings have implications for other railway staff, and staff in other industries in control roles with increasing automation.

7.3 How can temporal physiological data from wearable measures contribute to MWL assessment in rail industry live operations?

The answer to this first research question combines evidence from all three studies, both scoping reviews, and builds on existing research. The section focuses on how physiological data can contribute and its suitability as a MWL measure. It includes two novel contributions from the temporal EDA and HRV data: EDA can detect moments in workload; and average HRV has a strong negative correlation with average subjective workload ratings.

7.3.1 Moments in workload from Electrodermal Activity

EDA is a measure of the electrical conductance of the skin (Venables and Christie, 1980). This research progresses the previously very limited application of EDA to the railway signalling task. The research found EDA data can contribute to measuring MWL in two ways. Firstly, a key contribution of EDA is that SCR spikes can identify important ‘moments’ during a continuous task and secondly alertness and stress can be inferred from changes in EDA over time. EDA SCRs can indicate responses to sudden, unexpected events (Sokolov, 1963). The simulation study highlighted the importance of discrete periods, or 'moments' in workload. Such data could indicate periods of interest in the continuous data to better understand the individuals' experience of workload.

Whilst EDA was hypothesised to increase with phone calls (as found previously by Broekhoven 2016 in live signalling operations), this was not the case. The phone calls may not have been sufficiently novel or demanding to produce the stress response Broekhoven 2016) including participants knowing the simulator was not real (as noted in the perspectives and attitude study). In the simulation study, EDA SCRs identified points in time where participants appeared to experience a moment of realisation such as realising an error. These 'moments of realisation' occurred either at same time or with a delay after the external event. Tracking this time difference between external event and internal response could be a useful application of EDA. An example use would be to determine how long after an external event it takes an individual, such as a trainee, to notice an action is required, as implied by an increase in EDA.

This research suggests EDA indicates individual experience of workload. In the simulation study, EDA SCL identified times during the task where participants appeared to experience uncertainty, or time pressure. This would fit with previous research that has found that stress (Healey and Picard 2005) and alertness (Song et al., 2014) can be inferred from EDA. This research found no link between task demand and EDA, in contrast to previous research that suggested EDA reflects differences in task (Healey and Picard 2005), or changes in cognitive load (Mehler et al., 2012). In other words, it is not the number of trains alone that produce an EDA response, but instead an individual's experience of that task demand due to associated alertness and stress. Both stress and alertness were identified by participants as relevant to the signalling task. In this way EDA could provide visibility of individual experienced MWL that is not directly observable. This suggests EDA could contribute to a better understanding of the MWL experienced by signallers to show both brief moments and underlying level of alertness or stress.

7.3.2 Heart Rate Variability links with mental workload and confidence

The simulation study found that average HRV showed a strong negative correlation with average IWS (self-assessed workload), demonstrating that a high HRV was associated with low self-reported workload. This was a key contribution of the simulation study as limited research to date has applied HRV data to railway research. The correlation between HRV and IWS was no longer evident when values of HRV were adjusted by participants' baseline. This was an interesting finding as it suggested HRV was more indicative of differences between individuals' experiences of the task than it was between different levels of task demand. This is a novel finding, and one that differs from previous HRV research that found it was sensitive to task demand in pilots (Lehrer et al., 2010). If only average workload ratings were required, HRV could potentially replace IWS (self-assessed workload) ratings or be taken before a task to estimate an individuals' average self-assessed workload rating during a task.

HRV has also been found to correlate with emotional state, such as anxiety (Ramírez et al., 2015) which is traditionally considered to be a separate concept to MWL. A live trial with train drivers found HRV decreased before and after a train stop and at tunnels due to alertness and mental 'tension' (Song et al., 2014). Nickel and Nachreine (2003) proposes that HRV does not indicate MWL, but instead time pressure and emotional strain. HRV may imply uncertainty (Ramírez et al., 2015) with the simulation finding this as a trend rather than a significant difference between correct and uncertain observed behaviours. Participants in the perspectives and attitudes study indicated that detecting levels of confidence or uncertainty would be relevant to signalling, and time pressure was deemed less relevant. The findings of the simulation study, taken with the previous research suggest that low HRV shows promise as an indicator of individual high MWL. Ambiguity remains, however, around precisely what can be inferred from HRV data such as confidence, time pressure, emotional strain, or anxiety.

7.3.3 Timeline and storyboards of experienced MWL

Temporal data shows how physiological data changes over time relative to task events with the sequence of which changed first, plus rate of change and speed of recovery to baseline. The storyboard approach provided a way to visualise these changes based on the chronological data. The frequency breakdown used by previous research suits comparing discrete workload conditions in laboratory studies. The storyboard suit continuous data and are akin to the time sequences of train driver physiological data created during a live trial (Song et al., 2014).

The scoping review noted physiological measures were sensitive to different aspects of MWL (Matthews et al., (2014), Charles and Nixon (2019). This research used the storyboards to support a more detailed qualitative exploration of what specific aspects of MWL or task could explain why physiological data change. This includes preserving the interaction of the different timescales of changes including brief moments such as EDA SCRs that peak in seconds, compared to recovery to baseline that can take tens of minutes. In future, debriefs with staff in industry could use visualisations of temporal physiological data to discuss which changes in data values are associated with certain task events or MWL. If sequences are identified, they could ultimately lead to predicting MWL experienced and its likely impact on performance (positive or negative).

7.3.4 Cumulative impact of workload

A speculation regarding the physiological data that arose during the research was whether it could show a cumulative impact of workload. It is an idea that is based on the data in the simulation study that showed a large increase in Skin Conductance Level (SCL), particularly in participants P16, P17, and P20 (see Appendix C). These increases were not directly reflected in increases in subjective workload. It is postulated here that if one peak in EDA does not return to baseline before the next peak occurs (i.e., there is overlap in different instances of SNS activity), there will be a cumulative increase in SNS activity over time. In the literature on EDA, speed of increase is swifter than speed of recovery. In laboratory studies, recovery time is indicated by a recovery of EDA back to 50% of the peak value, relative to baseline. Theoretically therefore, in a continuous task, a series of separate events could lead to a cumulative, incremental, increase in sympathetic tonic baseline. A visualisation of this proposed cumulative increase is presented in Figure 7-1. The return to original baseline will be extended over time or fail to return to baseline during a continuous task.

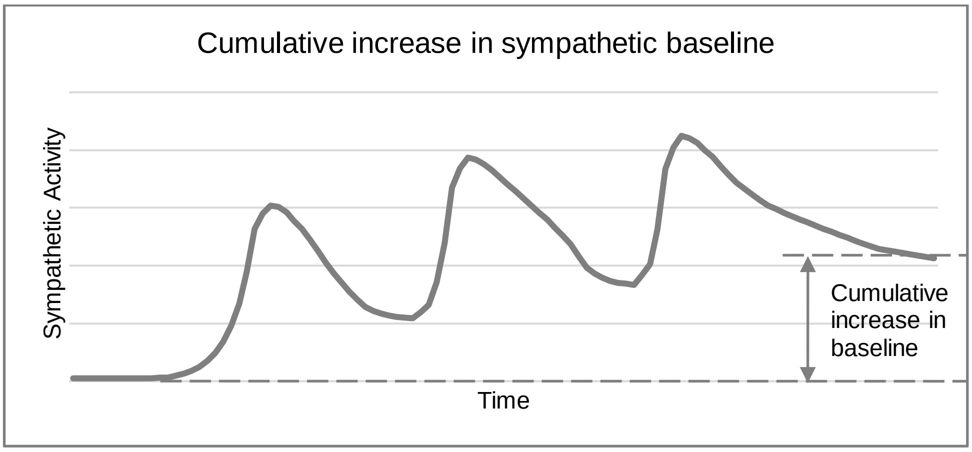


Figure 7-1 Cumulative increase in sympathetic baseline following multiple events

It is further proposed that if the individual notices the relative decrease in workload after one event, they may be unaware of any underlying overlaps and resultant increases in SNS activity. They may report a decrease in subjective workload back to the same scale point as before the increase, but their physiological state, in SNS activity, would be higher than before the increase. This fits conceptually with the theory around awareness of physical fatigue (St Clair Gibson 2018). Physical fatigue that is self-reported reflects the gap between an individuals' actual experience of fatigue relative to the individuals' anticipated level of fatigue. The implication is physical fatigue is not noticed if it is at expected levels. If this were to be found to be the case for cognitive effort, it would be interesting to explore when conscious subjective MWL matches underlying physiological state, and when the underlying physiological state shows an increase that the individual is unaware of.

Physiological data shows potential to, in future, detect and make visible when an individual is unknowingly moving towards an overloaded state. This type of data could provide visibility to managers, particularly those not in the room, of the effort or stress of staff. The managers could benefit from understanding the cumulative total hidden 'effort cost' of work on their staff of a shift, and compare that to the task demands, and operational performance.

7.3.5 Absolute versus relative values

The simulation study found both HRV absolute values and relative values offered different insight into MWL. It is common with physiological measures in research to use the relative value of physiological data (deduct participants' baseline) to reflect differences due to changes in task demand after removing individual differences. In the simulation study these relative values did not correlate with IWS (self-assessed workload). So relative HRV was not sensitive to task demand, which is contrary to previous research that did find HRV decreased when MWL increased (Lehrer et al., 2010).

Instead in the simulation study, the absolute value of average HRV showed a strong negative correlation with average IWS. This was one of the key contributions of the simulation study. If individuals' average HRV across the task was low, their individual average IWS was high. This suggests that the HRV physiological data is sensitive to individuals' subjective experience of workload rather than to changes in task demand. This would fit with individual's perceived load and internal load aspects of MWL proposed in Pickup's MWL framework (Pickup and Wilson 2007).

This thesis proposes that both absolute and relative values of physiological data are useful when understanding MWL. They answer different questions. Individuals' experience of MWL is shown in absolute values, whilst any changes in relative values indicate commonalities of response to MWL across participants (e.g. task demand). It should be noted that, in terms of MWL, the meaning of absolute values in physiological data are not currently well understood. Further research would be required to determine the ranges of absolute values to be expected during different tasks, and whether boundary values can be identified.

7.4 What are the theoretical implications of individual physiological data to changes in MWL in a workplace setting?

Drawing on findings from all three studies, and the literature reviewed, this research presents a preliminary conceptual model as its theoretical contribution. It proposes the measurement and monitoring of individual MWL through combining level of novelty task events with changing physiological state. The model comprises two scales that cover the concept of Novelty of Events and Autonomic State. Individuals are anticipated to move to the right over time as their breadth of experience increases. Individuals will remain in the upper half if they are rested and remain calm.

7.4.1 Novelty of events

The concept of Novelty of Events is a contribution this thesis makes to the theory and understanding of signaller MWL. MWL is a construct comprised of external task demands, internal individual experience of workload, and resulting task performance (Sharples 2019). Here Novelty of Events includes a sudden, or unexpected event. It also includes any new way of working such as due to the introduction of a new technology or a change in procedures. The novelty shows in physiological data if it causes surprise, a change in alertness level, or a stress response. The importance of the concept of Novelty of Events to the MWL of signallers, became increasingly apparent through the course of this research.

The initial industry interview study found within its results that drivers in their first year account for 25% of incidents (see 3.4.2 Incidents section) and the transition to novel technologies can be difficult for experienced staff (see 3.4.3 The simulation study confirmed that novel events may be inferred from EDA increases. The study went on to propose that uncertainty can be linked to novel events. Finally, the perspectives and attitudes study confirmed unexpected events happen frequently in signalling, so are a relevant element of the task to detect. The perspectives and attitudes study also concluded that novelty of events and uncertainty are inextricably linked, with confidence from experience a counterbalance to this. As novelty and unfamiliarity can introduce the risk of mistakes, the implications of novelty are worth further consideration.

Trainees and experienced signallers learn to deal with novelty in events through building experience over time, which in turn builds their confidence. This is drawn on concepts within existing literature such as expertise (Farrington-Darby and Wilson, 2006; Hoffman, 2014; Shanteau, 1992) and the skill-rule-knowledge framework (Rasmussen, 1983). The perspectives and attitudes study noted signallers’ confidence comes from having successfully dealt with certain situations on more than one occasion. Signallers, however, do not build up an even spread of experiences. Exposure to novel events can instead be individual, developing from the experiences an individual is exposed to (Bullough and Baughman, 1995).

Individual signallers' opportunity to gain experience of dealing with specific types of situations varies. Level crossings, for example, vary by the geography of workstations, being less common in urban rather than rural areas. In addition, signallers may go months without dealing with a particular type of situation if it occurs when they were not on shift. This means individual signallers will have gaps in their expertise. In live operations, novelty risk mitigations could include staff receiving assistance from colleagues and shift managers. Also tailored training could address gaps in experience such as simulator training or simple walk-through scenarios, to reduce the likelihood of a events being completely novel when it occurs during a real shift.

The use of the term 'Novelty of Events' draws on the literature and studies, linking the novelty of a given event and the resultant change in MWL for the individual. The concepts of sudden unexpected events and uncertainty were mentioned in the literature on physiological measures, but not specifically 'Novelty of Events'. The distinction between unexpected events and uncertainty relates to level of experience. In the simulation study this was seen in the data that neither the unknown timing of telephone calls nor arrival of the freight train produce an increased in EDA in all participants. Some participants instead showed increases in EDA, and a decrease in HRV, when they appeared uncertain.

A comparison can be made to the Skills Rules Knowledge levels of performance (Rasmussen and Jensen, 1974). Unexpected events, whilst unscheduled, may have been experienced before, allowing performance to occur at the rule-based level. A novel situation, with no pre-existing experience, requires a move to the knowledge-based level of performance.

The level of uncertainty experienced, at the rule or knowledge level, represents an awareness of the discrepancy between what the situation requires for successful performance and the likelihood of success based on individual experience.

7.4.2 Familiarity of events

Familiarity of Events is at the opposite end of the scale to Novelty of Events. As familiarity grows over repeated exposures, the size of EDA response reduces. In psychophysiology this is referred to as habituation. The scoping review of physiological measures found this is something that could be indicated by physiological data. The first few times a novel event is experienced, it would show as a spike in EDA SCR. As familiarity grows with that specific event, the spike would decrease in amplitude until the specific event was no longer novel and therefore occur in an absence of an EDA SCR spike (Hugdahl, 1995). This fits with the literature on experts who have confidence (Shanteau 1992) and can complete tasks with economy of effort (Hoffman et al., 1995). This aspect of EDA could be used to plot the progress of trainees as they develop from novices to experts. As individuals gain expertise, they are more confident (Shanteau 1992).

In the perspectives and attitudes study Signalling staff agreed with this, and that it was likely that a drop in stress would be seen in trainees as they gained experience. It should be noted however, that in the perspectives and attitudes study participants felt it would be unfair to monitor trainees or newly qualified signallers. This reflected a wish to avoid putting undue additional pressure on trainees. Another aspect of EDA to note is that, in laboratory studies, not all participants show EDA responses to the same stimuli (Boucsein, 2012, Braithwaite et al., 2013). As EDA can imply stress, it could be that non-responders are correctly showing an absence of stress response. To understand the EDA response, it would be important to collect data from both responders and non-responders for the same event.

7.4.3 Balanced Autonomic state

The Autonomic Nervous System (ANS) is comprised of two parts: the Sympathetic Nervous System (SNS); and Parasympathetic Nervous System (PNS) (Sherwood, 2013). Activation of the SNS increases sweating, Heart Rate, with an associated reduction in HRV, and inhibits digestion Table 4-1 Autonomic Nervous System). It prepares the body for 'fight or flight' emergencies. The PNS dominates in relaxed situations, stimulating digestion, and reducing the Heart Rate, with an associated increase in HRV.

The term 'Balanced Autonomic State' is used here to refer to a sustainable underlying physiological state that can support successful performance. This includes sustainable effort, concentration or vigilance, and alertness. Such a state is achieved with some sympathetic activity to raise alertness, but with variability to include periods of increased parasympathetic activity that allows the mind and body to rest, and for digestion during breaks. It is important to understand this is not about achieving and remaining at a specific point on a scale, but instead a variance that accommodates the needs of the human body. The Balanced Autonomic State includes 'flexibility' of state, so if physiological values increase, a swift return towards baseline follows. In terms of EDA, a Balanced Autonomic State can include EDA rises. The issue is having sufficient recovery time for EDA values to return towards baseline. If additional EDA rises follow in quick succession, the return to baseline will be delayed. The scale of EDA response varies both in individuals, as found in the simulation study, and across different work situations as found by Broekhoven (2016) with variance depending on the content and implications of different communications to signallers.

7.4.4 Skewed Autonomic state

A skewed autonomic state refers to a sustained period 'stuck' with either high sympathetic activity, or high parasympathetic activity. High sympathetic activity is associated with overload and stress (Sapolsky 1994), including due to novelty of events. High parasympathetic activity is associated with rest and lethargy, or reduced alertness due to underload. Either state will make successful performance more difficult to achieve. Such states could explain why individuals can become hyper fixated, with narrowed thinking in either during overload or underload. Physiological measures could provide a live read out of where on the sympathetic – parasympathetic balance individuals are. HRV would detect varying activity in the sympathetic and parasympathetic nervous system, EDA would detect only the sympathetic nervous system (Tortora and Derrickson, 2007). Measures would not aim to detect a specific value, instead variability centred around the personal baseline of an individual. If measures detect that an individual is moving away from their baseline then interventions could be implemented, such as allocating extra staff resource to help an individual or allowing the individual to take a break.

7.4.5 Novelty of Events and Autonomic State model

The Novelty of Events and Autonomic State (NEAS) model is a preliminary conceptual model (see Figure 7-2), drawing on findings from all three studies and the literature reviewed in the research. It proposes looking at MWL in a different way, from the perspective of changing physiological state and its relationship with human performance. It comprises two scales that cover the concept of Novelty of Events and Autonomic State. It considers an individual's level of expertise, likely speed of response and recovery of physiological state, the predicted level of EDA and HRV, and the effort required to maintain successful performance. Individuals will move along both scales during a shift, and over months as their experience and confidence increase. To illustrate the model, each quadrant is presented with individual's starting level of expertise and state, an operational scenario with likely performance, physiological response, potential human error, and resulting operational and individual impacts.

Quadrant I – Familiar Events Balanced State

The first quadrant combines Familiar Events with Balanced Autonomic State. It represents an individual who is highly experienced, with a wide range of operational experience, who is most likely to result in successful performance. This fits with research showing experts demonstrate a rich repertoire of strategies (Cellier et al., 1997). Experts are calm, confident in their decision making (Shanteau, 1992), and can make key operational or safety decisions. They have sufficient sleep and breaks to be alert, such as mid-way through a set of day shifts.

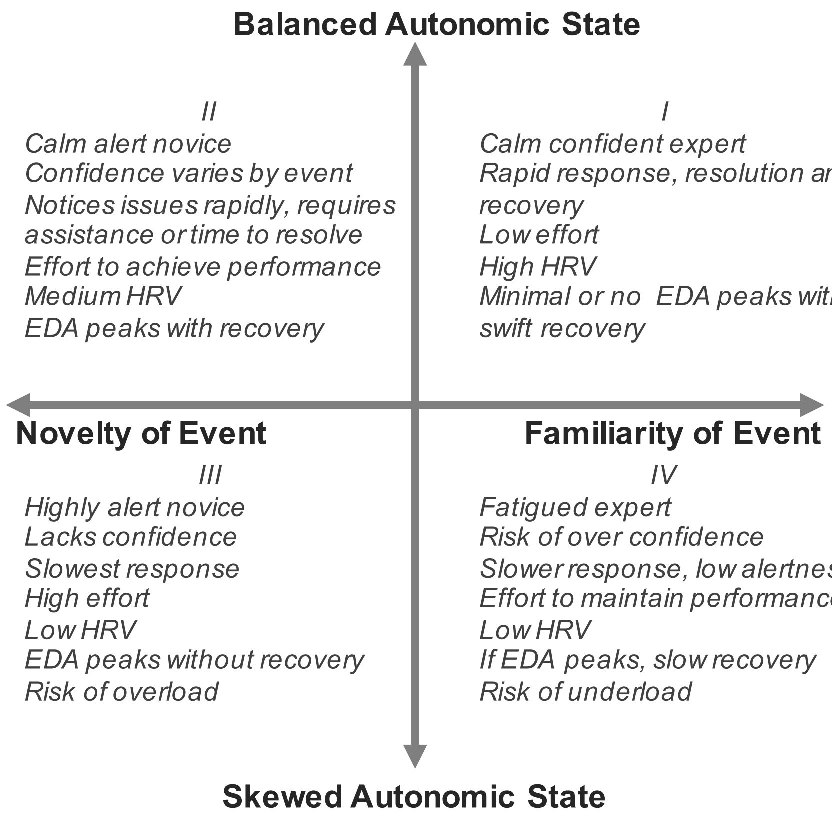


Figure 7-2 Novelty of Events and Autonomic State (NEAS) model quadrants

An example of an operational scenario in the top right Familiar Event Balanced State quadrant would be signalling trains during normal operations to meet the timetable. Normal operations include routine tasks such as: planning, setting routes; monitoring trains' progress; and operating CCTV level crossings (Balfe et al., 2008). It could also involve regulating trains by deciding which proceeds through a junction first. Verbal communications are minimal as operations are running normally, with signal aspect providing drivers their indication to stop or proceed.

This combination of high experience, balanced state, and a familiar task enable signallers to notice issues promptly and respond effectively to resolve them. It is the ideal combination for sustainable human performance, resulting in successful operational performance. Individuals will be able to achieve this without excessive effort as they have dealt with the type of event before. They will not be overloaded, underloaded, or under time pressure. Whilst they remain calm and work at a level they view as normal, or 'business as usual', they could seem busy to others despite being on top of the task.

In keeping with previous research on expertise, it is proposed individuals in this quadrant complete will their tasks with economy of effort (Hoffman et al., 1995), will be faster at problem solving (Chi et al., 1988) and adaptable. Individuals in this quadrant can adjust decisions continuously (Shanteau, 1992) to deal with disruptions swiftly and effectively to support the timetable. This includes the ability to deal effectively with rare or tough cases (Hoffman et al., 1995). An individual's underlying physiological state is predicted to have an average EDA SCL and exclude EDA peaks, as they have habituated to the events (Hugdahl, 1995), are not stressed, and their alertness is not excessively high. If something does occur that results in a peak in EDA, it would be a small and short duration SCR. Their HRV is anticipated to be high; reflecting confidence and the absence of uncertainty.

Errors of any kind are predicted to be at the lowest likelihood of all the quadrants. In this scenario there are no operational or individual impacts. Normal train running would continue. Successful performance could be sustained throughout a full shift.

Quadrant II – Novel Events Balanced State

The top left quadrant combines Novelty of Events with Balanced Autonomic State. It represents an individual who is a qualified signaller with a limited range of operational experience. They are calm but lack confidence in scenarios they have personally never dealt with before. They are rested and alert from sufficient sleep, rest breaks, and shift pattern, such as halfway through a set of day shifts.

An example of an operational scenario that would fit in quadrant II would be receiving a report of a cow on the line (as reported by signallers in the Perspectives and Attitudes study (see section Data relevance to signalling task). If this specific event is unfamiliar to a signaller, they can successfully apply the rules and procedures from their experience of other obstruction of the line incidents. The task involves stopping trains, arranging for staff to visit the site to locate the cow and block the point they gained access to the track, and then inform the driver to proceed at caution to confirm the track ahead is clear. Communication includes signalling and verbal communications to coordinate with staff to visit the site.

This combination of limited experience, balanced state, and novel task will enable an individual to be alert and notice issues. They may, however, respond slowly due to inexperience. It is a mixed combination for human performance that can be sustained, and result in successful task performance, but requires additional time or resources (such as support from colleagues). A decision made too rapidly increases the likelihood of error. Individuals will be able to achieve performance with moderate effort as they have not dealt with the event, or something similar, before. Whilst they can remain calm, they will require additional time or assistance to remain on top of the task and not become overloaded. The individual may seem alert but uncertain or hesitant.

Their underlying physiological state is predicted to have an average EDA SCL but include EDA SCR peaks in the situations most novel to them (Sokolov, 1963). These peaks may occur up to 14 seconds after the event they are associated with (Bound, 2016). HRV is predicted to be medium, as they are neither strongly confident nor uncertain.

Rule-based errors are likely, meaning the individual applies the wrong rule or set of procedures. Knowledge-based errors could occur in another scenario in this quadrant if the event does not clearly fit an existing rule or procedure. Individuals in this quadrant will benefit from support from colleagues who have relevant experience, such as the location's access points and farms in the immediate area. In this scenario, there are operational impacts but minimal impacts on the individual. There may be resulting delays to the route that need to be managed. They are likely to swiftly return to their normal state once the situation is resolved. In addition, the signaller gains valuable experience and confidence.

Quadrant III – Novel Events Skewed State

The bottom left quadrant combines Novelty of Events with Skewed Autonomic State. It represents an individual early in their career, with a limited operational experience. They lack confidence in situations they have not personally dealt with before. They experience strain from dealing with an event that is critical but novel to them, particularly if they are fatigued. This matches previous research that novices are less skilled than experts (Anderson, 2000). They are predicted to require more effort to complete tasks compared to the economy of effort of experts (Hoffman et al,. 1995).

An example of an operational scenario that would fit in the quadrant III would be an emergency call from a driver to say they may have hit something, as reported by signallers in the Perspectives and Attitudes study (see section Data relevance to signalling task), and in a study in live signalling operations

(Broekhoven et al., 2016). This type of event is sudden and unexpected. To deal with it, the signaller must speak to the driver to both reassure them and to gain a report of the event. In addition to speaking to the driver, they must signal other trains to avoid the stopped train and limit the disruption. The need for verbal communication is considerable, not only to the driver but with colleagues, to coordinate with multiple agencies. Signallers in the attitudes study added that drivers only call when there is a problem, so just hearing the phone ring can be stressful, particularly until the nature of the incident is understood.

This combination of limited experience, skewed state, and novel task limits an individual’s ability to notice issues. It will increase the time issues go unnoticed, and the time needed to resolve them. It is the most challenging combination for human performance, resulting in reduced sustainable duration and risks of poor performance outcomes. Individuals will need to employ excessive effort to compensate for inexperience and lack of confidence. They are likely to seem uncertain. They are at risk of overload and strain.

Their underlying physiological state is predicted to include frequent EDA SCR peaks, with elevated EDA SCL. This matches the findings of a study in live operations, where a signallers experienced more frequent EDA SCRs and their EDA SCL was significantly higher during periods of high self-assessed workload (Broekhoven et al., 2016). Their high workload took over forty minutes to lower when they received a call from a driver reporting they may have hit a person (Broekhoven et al., 2016). HRV, in comparison, is anticipated to be low due to their uncertainty and lack of confidence.

This quadrant comes with the greatest risk of all types of error. Individuals in this quadrant will benefit from both support from colleagues and regular breaks. It would take time for the individual signaller’s state to return to baseline. Both could extend past the end of the shift.

Quadrant IV – Familiar Events Skewed State

The bottom right quadrant combines Familiarity of Events with Skewed Autonomic State. It represents an individual who is highly experienced, with a wide range of operational experience. They are confident but are likely to experience vigilance decrement from cumulative fatigue, such as at the end of a series of night shifts.

An example of an operational scenario that would fit in the bottom right quadrant IV would be dealing with a disrupted timetable, which occurs frequently enough to be familiar to the individual signaller. To recover from the disruption, tasks include regulating trains, by determining which proceed first through a junction, weighing the level of delay of each train. The need for verbal communication is low, as signal aspect provides drivers the indication they need.

This combination of high experience, skewed state, and familiar task will mean individuals may be less alert and slower to notice issues or respond to resolve them. It is a mixed combination for human performance that can result in successful task performance but will require rest and sleep as it cannot be sustained. Individuals can achieve successful performance with additional effort to boost alertness and remain on top of the task. There is a risk of over-confidence here, if an individual fails to realise they are fatigued. Individuals are bad at judging their own fatigued state (Martindale, 2012). If the event is very familiar, then underload is a risk. The individual may seem subdued and unable to maintain their attention.

Their underlying physiological state will have low EDA SCL. Any EDA SCR peaks are predicted to take longer to recover to baseline. HRV is anticipated to be low, with increased uncertainty and effort.

Skill-based errors of slips and lapses are likely in this quadrant. An example, during the scenario presented above, would be the signaller mistakes two train head codes (the train reporting numbers), wrongly routing the first train onto a route intended for the second train. This type of error is likely to lead to further operational impact if the driver takes the route, they will miss a scheduled stop, disrupting passengers. If the driver notices the wrong route they will stop and call the signaller causing a slight delay. If the signaller notices their error, they may choose to change the signal to red, however this risks the driver applying the emergency break which could injure passengers. If a signaller did wrongly route a train, it is proposed the individual would experience a physiological response when they realise their error. In this scenario there are individual impacts with potential operational impacts. Individuals would benefit from a break or ideally sleep, to improve their performance. Their vigilance decrement could lead to an error that impacts operational performance.

7.4.6 Comparison of NEAS model with existing MWL models

The Novelty of Events Autonomic State (NEAS) model matches elements of existing models of MWL to varying extents. Comparisons are made here with deWaard's workload and performance model (deWaard 1996), Xie and Salvendy's factors contributing to mental workload (Xie and Salvendy 2000), and Edward's precipice of performance (Edwards et al., 2016).

In their model of workload and performance, deWaard proposed how these have an inverted U relationship with each other as demand increases (see Figure 2-9). Here it is proposed that deWaard's regions from A to D map onto the NEAS model anticlockwise as shown in Figure 7-3.

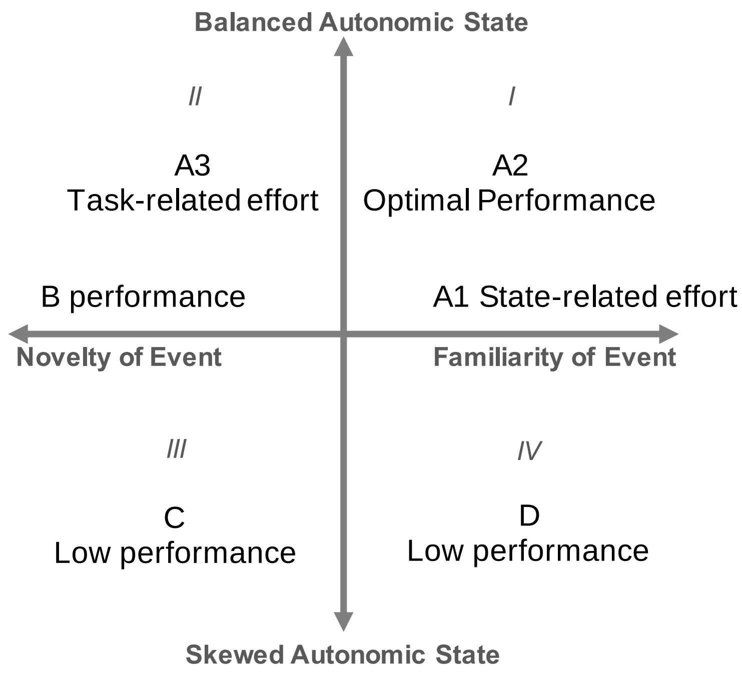


Figure 7-3 Mapping deWaard's MWL regions A – D onto the NEAS model

DeWaard's state-related effort (region A1) sits on the boundary between quadrants IV and I. Here an increase in state-related effort is needed to sustain performance. Optimal performance (region A2) maps onto quadrant I Familiar Events Balanced State. Task-related effort (A3) maps onto the quadrant II Novel Events Balanced State. In this area, performance is sustained, with effort but without overload. Region B, when task-related effort can no longer sustain performance, maps onto the boundary between quadrants II and III. Region C, where performance is low, maps onto quadrant III Novel Events Skewed State. Finally, region D maps onto quadrant IV Familiar Events Skewed State. In this case, the NEAS model takes existing theory and builds on it to propose what measurable physiological changes could be detected as an individual transitions between the quadrants and through the regions of deWaard's model.

Edwards' 'precipice of performance' (Edwards et al., 2016) fits the same position on the NEAS model as deWaard's Region B; on the boundary between top left and bottom left Novel Events - Balanced State and Skewed State (see Figure 7-3). In Edward's theory (see Figure 2-10), performance can be protected by compensatory strategies (which would map onto the Novel Events Balanced State quadrant). After these compensatory strategies fail, performance degradation is rapid, resulting in a state that fits quadrant III Novel Events Skewed State. After the rapid degradation in performance, Edward's describes the individual as having no plan, likely to panic, and reliant on colleagues' support.

A key difference between what is proposed by the NEAS model compared to both deWaard's and Edwards' theories is the concept of a quantity of demand is absent. Instead, the model suggests a quality in the concept of novelty. Whilst quantity of task demand is a core concept in MWL, in this research physiological measures show mixed results when used to detect task demand. Previous research comparing physiological measures for detecting MWL have found mixed results (Charles and Nixon, 2019; Matthews et al., 2014). Specific to HRV and EDA considered in this research, HRV decreases with increased task demand (Lehrer et al., 2010) and EDA with cognitive load (Mehler et al., 2012). Equally, research found EDA changed with task difficulty (Healey and Picard, 2005) and arousal (Song et al., 2014). In the simulation study within this thesis, HRV and EDA did not correlate with Task Demand, but average HRV did correlate with average self-assessed workload ratings. It is proposed by the thesis, therefore, that novelty of events has as great an impact on an individuals' capacity to deal with workload as the quantity of task demand.

A subset of factors that contribute to mental workload in a complex setting (Xie and Salvendy, 2000) map onto the NEAS model. Fatigue (Klein and Malzahn, 1991), task duration (Dember et al., 1993), and stress (Hart and Staveland, 1988) will move an individual down in the model. Knowledge will move them right, and task uncertainty (Knight and Salvendy, 1981; Lehto and Buck, 1988) will move them left. It is proposed here that task importance (Sawin and Scerbo, 1993) may move them down in the model, especially if it leads to strain. The concepts of attitude, motivation (Reid and Nygren, 1988) and task complexity (Hart and Staveland, 1988) do not directly map onto the model. Whilst these are accepted as factors that can influence workload and performance, they are not factors which physiological measures are likely to delineate.

Taken together, the above comparisons highlight the challenges of determining the diagnosticity of physiological measures for MWL, when the construct of MWL is not well defined. The models that map more closely with the NEAS model are those that distinguish the individual factors such as fatigue (Klein and Malzahn 1991), and performance outcome, rather than those focused on external task factors such as task complexity (Hart and Staveland 1988). This presents an opportunity to re-evaluate what we mean by MWL.

7.5 Progressing the terminology around MWL

This research found that individual experience of workload can be inferred from physiological data. Alertness, stress, and uncertainty were deemed the aspects of workload most relevant to signaller workload. Increases in these three aspects of workload can be inferred from increases in EDA and decreases in HRV. It could be concluded from this that EDA and HRV can be used to measure MWL. This, however, depends on the definition of MWL. Suitability of workload measures criteria includes sensitivity to changes in task demand and validity to detect only cognitive aspects of MWL (Eggemeier et al., 1991, Sharples and Megaw 2015). Based on these criteria EDA and HRV do not measure MWL. What this research proposes instead is that physiological data can assist in determining why individuals respond to the same task demand in different ways. In this respect it could be argued that physiological data could assist in distinguishing reasons for change that are individual in nature. In this way EDA and HRV could provide diagnosticity.

Accurate use of terminology will be important to remain clear on what is detected by physiological measures (e.g. HRV is the time gap in electrical signal to the heart) and what can be inferred (e.g. high experienced workload when HRV is low). This research proposes this clarity will assist in incorporating these new measures and their data into our existing understanding of MWL. The temporal data from measures can provide information on the sequence of events in task, and any time gap between external events and internal reaction. In combining our existing understanding of what external task demands affect MWL, physiological data can assist in building our understanding of the internal demands and reactions to the task including effort, strain, and uncertainty. This goes beyond assessing workload as a quantity of task demand and includes the combination of factors that support successful performance.

7.6 What are staff perspectives on wearables and use of their personal physiological data?

Staff attitudes and opinions on the use of wearable measures are an important factor to their success. The devices were not viewed as a status symbol, unlike new technology in the TAM2 model (Venkatesh et al., 2000). This could be due to wearables not directly supporting performance outcomes and not being viewed as a privilege. As wearing measures are novel to all staff for use at work, a certain level of hesitance or resistance was expected. Overall response by participants in the perspectives and attitudes study was that they would be prepared to trial the measures. They were unsure what the reaction of colleagues would be.

During the simulation and perspectives and attitudes studies, a distinction developed between the suitability of wearable devices for the live environment and the use of data from wearables.

7.6.1 Device suitability for use in live operations

The perspectives and attitudes study indicated that wearable devices would potentially be appropriate for use in the live operational environment. The wrist strap received acceptable comfort ratings, with minimal distraction. It was predicted that the acceptable wearable time could be up to a full shift (8-12 hours). Some participants commented that the wrist strap would be like wearing a watch, so they would be unaware of wearing it. In comparison, participants were uncertain about wearing the chest strap at work, with a wider range of opinions being offered about the chest strap. No participants currently used a chest strap as a fitness tracker. It could be this preference for the wrist strap, in part, reflects the relative difference in familiarity and positive experiences of a wristwatch compared to the novelty of wearing a chest strap. This fits with the TAM2 model of past direct experience of technology influencing attitudes (Jacobs et al., 2019; Venkatesh and Davis, 2000).

Comfort ratings were not collected during the two field studies that applied wearable measures in live railway operations (Song et al., 2014; Broekhoven et al., 2016). Based on their data collection, drivers wore a chest and a finger EDA sensor for up to 3 hours (Song et al., 2014) and signallers wore a wrist strap for an average of 3.4 hours (Broekhoven, 2016). Taken together the implications of these findings are that a future live trial of wearables could go ahead, with the wrist strap being the preferred device and the chest strap optional if a period of familiarisation is provided for participants to get used to wearing it.

7.6.2 Suitability of personal physiological data use

The perspectives and attitudes study found the concerns of staff were less about the use of devices to collect the data and more around the consequence of use of their data. The study determined that trust was a key factor. It was essential staff had trust in those who would see their data, particularly any named data. Whilst they accepted researchers and third-party providers would need to see the data, their concern was mainly whether their managers would see data that could be linked to them. They were concerned that physiological data may be misused to assess their competency. Their ultimate concern was that they may lose their job. Concern was highest when those who saw their data had the most direct control over their employment, and the perceived risk that managers may use the data against the staff. Where data from the wearables could be used to support staff, perspectives and attitudes to their use were improved.

The above reflect organisational cultural barriers around the implementation of wearable physiological measures in the railway industry. Such barriers are not restricted to physiological measures. Similar concerns and cautiousness are evident around fatigue management, as reported in Chapter 2. Examples include staff fearing reporting fatigue, and there being contradiction between shift patterns designed to reduce fatigue risk and management agreement to shift swapping. In this organisational cultural context, it is understandable that staff would wish to have clear justification prior to the use of personal physiological data and agreement as to how their data would be used.

Anonymity or aggregation of results may, in theory, increase acceptance. A live trial to collect anonymized physiological data could proceed to prove what it can show. In a signalling centre anonymisation is difficult however, as individuals can be identified from the roster by date, time, and workstation. Instead, this research recommends improving perceived usefulness of the measures to improve attitudes. To achieve this, the industry would need to be clear what operational benefit or improvement would come from physiological data. An example would be informing shift break patterns based on alertness levels, or tailoring training scenarios from physiological responses. This would alleviate a dichotomy from the findings that absolute individual physiological data is a useful contributor to understanding individual workload but difficult to anonymise. Relative data, with baseline removed, would be easier to anonymise but may not show individual experience of workload. A tangible benefit to railway operations would be to encourage staff to accept wearable measures including use of absolute, identifiable data, rather than only anonymized data.

7.7 Limitations

Regarding participants, railway signallers became the focus. The potential exists for the results here to have implications for other staff roles in rail (e.g. drivers) and for control staff in other transport industries. To compare perspectives across different transport industries, a broader range of stakeholders could be recruited. This could be an interesting study, particularly to understand the relative difference in underlying organisational cultural maturity regarding potential acceptance of physiological measures.

The simulation study found individuals differ in their physiological data. To develop physiological measures for MWL assessment it will be important to understand the impact of these individual differences. A repeated measures study, with data collected from the same individuals, would assist in understanding how stable (and therefore predictable) individuals' responses are over time and the reliability of the measures.

The simulation study had a higher proportion of EDA non-responders than previous studies (Braithwaite et al., 2013). This could be caused by sensor placement, low room temperature, or it could be correctly showing an absence of stress response to the task. The latter reason could be an indicator of positive staff state. To understand the EDA response, it would be important for future studies to collect data from both responders and non-responders for the same event, rather than remove non-responders. Linked with the above recommendation to repeat measures, it could be determined whether certain individuals are more likely to be consistent non-responders.

7.8 Implications for industry

There are a range of implications for the railway industry relating to the use of wearables, what benefit the physiological data could provide industry, and how to progress these. Overall physiological wearable devices could suit use in live operational environment, particularly the wrist strap. Physiological data was found to reflect individuals' experience of MWL rather than task demand, as found by some previous research. Temporal physiological data could make individual effort more visible and provide useful feedback on staff individual MWL. This would be a form of mutual monitoring, currently achieved informally in operations through teams observing one another and stepping in if required such as answering a phone when a colleague is busy.

Physiological data could not answer the question of how many trains are too many for one signaller to handle. Physiological data could answer which situations do signallers find the most challenging. EDA, collected using the wrist strap, could identify moments of increased workload during a continuous task in live operations, including moments of uncertainty, time pressure, or realisation such as an error. This includes moments they were conscious of, and those they may not have been consciously aware of. HRV, collected from a chest strap, could indicate average individual workload. If only average workload rating were required, average HRV could replace self-assessed workload measures. The acceptance of a chest strap was lower than a wrist strap, so would require a period of familiarisation. The benefits of physiological measures would be to highlight individual experience of workload, and their unseen effort, level of alertness or strain required to complete a task. This also presents the future potential to monitor fatigue, to fit with existing programs that manage shift work and the risk of fatigue.

Physiological data could produce storyboards of a continuous task, with other data sources to indicate task demand, phone calls, other events in the task, shift pattern and operational performance status. These could facilitate debriefs with staff to understand periods of both sustainable effort with successful performance and any challenges. The data would need to remain anonymous whilst progressing these measures from the laboratory to the live environment. Further research is required to confirm which specific data from EDA and HRV is most relevant to the signalling task. This could then inform industry agreement between staff and managers on who can see the data and how it would be used. Ideally such data could inform tailored training sessions (using walkthroughs or simulator) to fill individual gaps in work experience and build confidence in situations that would be challenging. The aim would be, ultimately, to make the workload experience and effort of staff visible to managers and trainers.

The use of physiological measures is not appropriate in the railway industry yet. Prior to implementing live use of physiological measures, it is essential for staff to accept their use as the measures rely on staff personal data. To do this involves a tangible benefit to railway operations and demonstrate how the data could be used to support rather than blame individuals for their performance. A way to progress this would be to develop the NEAS model with staff and managers to determine how best the workload of staff can be managed to support successful operational performance.

The benefit of applying the NEAS model would be to understand where an individual sits to identify how best to support them and maximise the time spent in the top right FB region (through support and rest). Staff and managers would benefit from understanding the distinction between how rest breaks could be used to recover from a Skewed Autonomic State. To mitigate the risk of uncertainty and novelty of events involves providing staff support from experience staff or providing additional time for novel situations, build individuals' experience and confidence, and retain staff with experience. This works for stage of career from staff new to signalling to those transitioning to using more modern signalling equipment. The aim is to build familiarity and confidence in individuals to successfully deal with a wide range of situations or events.

There are broader implications for industry of the findings of this research. Firstly, novelty of events has been identified as a source of increased workload and effort. Building experience and confidence in staff can mitigate the risks associated with this novelty, as stated above. In addition, industry could predict when changes in technology or procedure introduce an element of Novelty of Events. This should be expected to have a negative impact on performance, for a period, until whatever has changed becomes familiar.

Secondly managers of staff would benefit from understanding what tailored support they can provide staff to build their experience and confidence. Part of this is to understand the position of an individual within the NEAS model is not a fixed point. It will move gradually to the right over time, as experience and confidence grow, whilst introducing novelty will move them left. During any 24-hour cycle everyone will go through periods where they drop lower, that can be managed by shift patterns and break patterns. The advantage of this approach is to demonstrate the benefit of understanding physiological state of individuals and how that understanding can help staff to sustain their efforts and in turn support successful operational performance outcomes.

Chapter 8: Conclusions and future work

8.1 Chapter overview

This chapter completes the thesis by presenting the conclusions of the research. This includes the contributions of how physiological measures can contribute to MWL assessment, staff perspectives and attitudes on their use, and theoretical implications. The chapter finishes with recommendations further research and implications for future work application in the rail industry.

8.2 Conclusions

This research contributes to our understanding of human performance in the rail industry by determining the potential contribution of temporal physiological data from wearable measures. The research addresses three research questions:

1. How can temporal physiological data from wearable measures contribute to MWL assessment in rail industry live operations?
2. What are the theoretical implications of individual physiological data to changes in MWL in a workplace setting?
3. What are staff perspectives on wearables and use of their personal physiological data?

The specific contributions of the research are how temporal physiological EDA and HRV data could contribute to MWL assessment of railway signallers, whether wearable measures suit use in live operations, and what staff perspectives are on the use of their personal data. From these the research also proposes theoretical contributions around MWL and the implications of these for the rail industry.

The rail industry is challenged by staff overload and underload as they negatively affect human cognitive performance. This is particularly the case during disruption and is despite increasing automation. To address these MWL related challenges, the PhD considered ‘what does good look like’, rather than focusing only on errors. This includes what range of MWL supports successful operational performance, and what MWL is sustainable by staff. As the research progressed it focused on signalling staff, although the results have implications for other control staff both within the rail industry and in other industries requiring staff to complete safety critical tasks. Physiological wearables were identified as a potential measure of MWL that suit use in live operations. Using wearables to measure MWL is relatively new, particularly in an industrial field setting. To date both EDA and HRV have mainly been constrained to laboratory and simulator settings. To address the first research question, the PhD investigated what EDA and HRV data could contribute to MWL assessment.

The first research question was ‘how can temporal physiological data from wearable measures contribute to MWL assessment in rail industry live operations?’ In answer to this, regarding temporal physiological data, the research found that both EDA and HRV data could contribute by indicating individual experience of workload. EDA measures skin conductance. The novel contribution regarding temporal EDA data was that EDA SCR spikes can identify important ‘moments in workload’ from a continuous task. These may be sudden unexpected events or a moment of realisation, or uncertainty, when an individual realises there is a problem. Such data could provide information for a debrief to staff and tailor training to address events individuals find challenging. In addition, EDA SCL, the underlying baseline value of EDA, could imply levels of uncertainty or alertness. The research indicated that changes in stress and alertness, were due to either unexpected, or novel external events such as incidents. In future, EDA could monitor alertness, with a decreased SCL indicating reduced alertness. Equally, if EDA

SCL remained steady, an absence of SCRs could be a positive indication that an individual is confident and not under strain.

HRV shows promise as an indicator of individual MWL. The novel contribution of HRV was that average HRV values across a task had a strong negative correlation with average subjective workload ratings. This means that individuals with higher average subjective workload have lower HRV. Previous research suggests confidence, time pressure, emotional strain, or anxiety can be inferred from HRV. Further research is required to confirm precisely what aspect of MWL HRV is sensitive to. An interesting and novel find of the research was that HRV was more indicative of differences between individuals' experiences of the task than it was between different levels of task demand. The correlation between HRV and IWS was no longer evident with relative values of HRV, with participants' baseline removed. In future, if only average MWL was required, HRV could potentially replace IWS as a self-report workload measure. This would provide an indication of how individuals varied in how much confidence versus strain they experienced during the task.

This research found physiological data did not indicate the number of trains an individual was dealing with but instead an individual's experience of that task demand due to associated alertness and stress. In this way physiological data could visualise the individual experienced MWL that is not directly observable. In future physiological data could be used to provide storyboards for use in debriefs with staff. Visualisation of the temporal data could provide a basis for discussion to understand what task or individual factors influenced individuals' workload. The benefit would be to understand periods of both sustainable effort with successful performance, and events or periods of uncertainty that were challenging. In turn, this could inform tailored training for staff to practice events individuals found challenging, to build their confidence. If data were collected real-time, it could provide visibility to managers, particularly those not in the room, of the effort or stress of staff.

The managers could benefit from understanding the cumulative ‘effort cost’ of work on their staff. Ultimately, if sequences were identified, this could lead to predicting when staff MWL levels are at risks of moving from successful performance into the higher risk areas of underload or overload.

In answer to the first research question, regarding use of wearables in live operations, the research determined wearable devices could suit use in live operations. Staff supported the idea of using wearables to improve the measurement of MWL in live operations. In particular, the wrist strap was identified as suiting live operations as it was viewed as more comfortable and distraction due to the device would be minimal. In addition, staff reported that inferring alertness from the EDA data collected from the wrist strap would be relevant to signallers. The chest strap was unfamiliar to staff, compared to wearing something on the wrist. This suggests a period of familiarity would be essential if a chest strap was required. HRV data from the chest strap could indicate how confident or strained an individual is, or how they perceive their MWL. Over time, if chest strap wearables become more prevalent for personal use, familiarity can inform staff acceptance of use in a work setting. It is important the industry acknowledge that, as with the technologies being implemented for direct task support, the acceptance of new technologies takes time.

The second research question was ‘what are the theoretical implications of individual physiological data to changes in MWL in a workplace setting?’ In answer to this, the theoretical contribution of the research is the preliminary conceptualisation NEAS Model. This provides a way of understanding how Novelty of Events and Autonomic State combine to impact individual MWL and performance outcomes in a quadrant. It suggests a way of understanding how individual’s physiological state combines with novelty of event to predict likely performance. The model considers across two axes of physiological state and level of confidence relating to how novel an event is that needs staff input. Physiological data provides a potential way to identify where in the NEAS model an individual is, and what would support an improvement in performance. This contribution progresses the theoretical understanding held around individual experience of workload and how it has an important contribution to make in the measurement and prediction of MWL.

The physiological data in this research, and the NEAS model, suggest a new way to consider MWL of staff in live operations. Compared to previous MWL theories, the concept of a quantity of demand is absent. Instead, the model suggests a quality in the concept of novelty. Traditionally the concept of MWL does not include emotional state. It focuses instead on quantity of task demand and of cognitive information processing. Whether EDA and HRV are measures of MWL depends on definition of MWL. Suitability criteria suggest measures are sensitive to task demand and not emotions. Signallers indicated however that detecting levels of confidence or uncertainty would be relevant to signalling. This research proposes instead that physiological data can assist in determining why individuals respond to the same task demand in different ways. In this respect it could be argued that physiological data could assist in distinguishing individual reasons for change in the data. In this way EDA and HRV could provide diagnosticity of MWL.

The third research question was 'what are staff perspectives on wearables and use of their personal physiological data? In answer to this, the research identified that concerns around the use of wearable measures related more to the potential consequences of use of personal physiological data. Trust was a key factor. Staff were hesitant as they thought their data may be used to assess their competence. Their ultimate concern was that they may lose their job. One way of addressing this would be to anonymise the data, however this would lose much of the potential value of the data in understanding individual factors. The PhD recommends the industry instead aims to use the physiological measures only when the staff are prepared for named data to be shared. This is to ensure both the individual staff member, and those with responsibility to support them, can benefit from the data. In the meantime, industry can benefit from lessons learnt from the theoretical implications identified by the research findings.

8.3 Future work

The use of physiological data shows potential to provide valuable insight and visibility of individuals' experience of MWL. Further research would help progress the use of physiological measures. Research should continue to explore what specific aspects of MWL can be inferred from specific physiological data. This would clarify what individual MWL factors the physiological data are sensitive to such as time pressure, stress, alertness. This research indicates physiological data are more sensitive to individual factors, so it is recommended that they are not used in studies that only track quantitative Task Demand.

It would be interesting in future to confirm whether individual MWL is cumulative and in what circumstances individuals' awareness of their MWL matches their physiological data. The research found that average HRV negatively correlated with average self-assessed workload. Previous research, however, suggests that individuals may be more aware when there is a discrepancy to their anticipated levels. Combining these two, it would be interesting to investigate when an individual does not report high workload as it matches their anticipated MWL level. Physiological data could provide visibility of their underlying state. Equally, if MWL is cumulative, individuals may be unaware their underlying state has incrementally changed over time. Again, physiological data could indicate when data returns to baseline (indicating it is physiologically more sustainable), and when it does not return to baseline and individuals may be unaware.

This research recommends both absolute and relative values of physiological data are considered in future studies. Both can be useful to understanding

MWL. They answer different questions. Relative values, adjusted by baseline, provide information on task differences in MWL as it removes individual differences. Absolute values retain the individual differences in experience, which are deemed a strength of the measures rather than something that should be removed. Future studies applying a repeated measures approach would also be desirable, to determine the reliability of measures both within and across individuals.

The outcome of the research for industry is that wearable physiological measures could benefit the rail industry in future but should not be used yet. Cultural barriers within industry would need to be addressed prior to use. Regarding the cultural barriers to acceptance, resistance to the measures mainly relate to concerns around the consequence of use of personal physiological data. The ultimate concern of staff was that they may lose their job. Trust was identified as a key factor, with staff hesitant to provide data that managers may be used to assess their competency. Instead of introducing new physiological measures, this research recommends applying strategies to support staff remain alert and confident. This could include effective use of breaks and building confidence through experience. The NEAS preliminary conceptualisation could assist staff and managers to understand how individuals, over time, are affected by novelty and would benefit from different support. Retaining staff with experience and building staff confidence can mitigate the risks associated with novelty of events. Breaks can mitigate the risks of fatigue and reduced alertness. These strategies would not require staff to wear physiological measures. Instead, if effective, they would provide a justification for why monitoring individual state could be mutually beneficial to staff and managers in future.

If a justifiable purpose is determined, with clear operational benefit, then a trial of wearables could be run in live operations, initially using only the wrist strap and EDA data. The chest strap was less familiar to staff and viewed as less comfortable. A trial would provide staff the opportunity to experience wearing a device and to validate the data. A debrief with staff after the trial could identify what aspects of the task, experience, and effort the data shows to evaluate the utility of the process. Additional data about task demand including phone calls, level crossings, and control inputs would be required to compare physiological responses. The initial trial would need to ensure staff data was anonymised. The trial could demonstrate to staff the benefits of sharing graphical data, such as to show the physiological effort required to sustain successful performance. This data, if shared, could inform tailored training. The physiological data shows the potential to, in future, detect that an individual is moving away from their baseline so timely interventions could be implemented.

Appropriate choice of physiological measures depends on what question needs answering. The strength of physiological measures is to infer individual experienced workload, rather than task demand. If the question is 'what quantity of task demand can staff handle?' This cannot be answered by physiological measures. If the question is 'how can we visualise individual workload changes over time?' then temporal physiological data from wearables could provide an answer. The latter would be choosing physiological measures for their strength of detecting individual experience of workload. The data could show the effort that may not be directly observable and do so with minimal task interference.

Whilst the results focus on railway signallers, these findings have implications for other railway staff, and staff in other industries in control roles with increasing automation. This research opens the opportunity for similar research to be applied in live operations in other industries were managing individual workload and strain is a protective factor for overall sustainable safe operations.

The research concludes that temporal physiological data shows great potential to contribute to the MWL assessment of signallers. HRV can indicate self-assessed workload, as it shows a strong negative correlation with self-assessed workload. EDA could identify moments of realisation during a continuous task to support debriefs with staff and inform training needs. Uncertainly or confidence could potentially be indicated by both HRV and EDA. Further research is required to clarify what specific aspects of MWL different physiological data are sensitive to. Collecting physiological data from staff should, however, not happen yet. Whilst wearable measures could suit use in live operations, staff concerns around how data will be used need to be address first. In future temporal data could indicate when staff are at risk of moving from good performance, into the higher risk areas of either underload or overload. In the meantime, the theoretical contributions of the thesis can benefit industry practice. Increased awareness of the impact of physiological state and novelty of events can inform effective MWL management. This in turn can support sustainable staff effort to achieve successful human performance in live operations.

References

Åkerstedt, T., 1991. 'Shift Work and Sleep Disturbances' in Peter, J.H., Penzel, T., Podszus, T., von Wchert, P., (eds.) *Sleep and Health Risk*. Berlin, Heidelberg, Springer, pp.265-278. doi.org/10.1007/978-3-642-76034-1_32

Anderson, J.R., 2000. *Cognitive Psychology and its Implications*. 5th Edition. New York: Worth Publishers.

Anund, A., Fors, C., Kecklud, G., van Leeuwan, W., Åkerstedt, T.J., 2015. *Countermeasures for fatigue in transportation: a review of existing methods for drivers on road, rail, sea and in aviation.* Swedish National Road and Transport Research Institute (VTI). https://www.diva-portal.org/smash/get/diva2:807456/FULLTEXT01.pdf [Accessed 27.10.2022]

Ashton, R. and Fowler, A., 2005. 'Human friendly rosters: reducing the risk of fatigue' in Wilson, J.R, Norris, B., Clarke, T., and Mills A. (eds.) *Rail Human Factors.* Abingdon, UK: Routledge, pp.203-214.

Averty, P., Athenes, S., Collet, C., Dittmar, A., 2003. 'Evaluating a new index of mental workload in real ATC situation using psychophysiological measures' in *The 21st Digital Avionics Systems Conference, IEEE*, pp. 7A4-1-7A4-13. doi.org/10.1109/dasc.2002.1052916

Ayaz, H., Shewokis, P.A., Bunce, S., Izzetoglu, K., Willems, B., Onaral, B., 2012. Optical brain monitoring for operator training and mental workload assessment. *Neuroimage* 59, pp.36–47. doi.org/10.1016/j.neuroimage.2011.06.023

Ayaz, H., Willems, B., Bunce, S., Shewokis, P. a, Izzetoglu, M., Hah, S., Deshmukh, A., Onaral, B., Bunce, B., Shewokis, P. a, Izzetoglu, K., Hah, S., Deshmukh, A., Onaral, B., 2010. 'Cognitive Workload Assessment of Air Traffic Controllers Using Optical Brain Imaging Sensors' in Marek, T., Karwowski, W., Rice, V. (eds.) *Advances in understanding human performance: Neuroergonomics, human factors designm and special populations*, pp.21–31. doi.org/10.1201/EBK1439835012

Baddeley, A.D., 2001. Is working memory still working? *American Psychologist,* 56, pp.851–864. doi.org/10.1037/0003-066X.56.11.851

Bainbridge, L., 1983. Ironies of automation. *Automatica* 19(6):775-779. doi.org/10.1016/0005-1098(83)90046-8

Baker, C., 1962. *Man and radar displays.* Oxford: Pergamon Press.

Balfe, N., Crowley, K., Smith, B., Longo, L., 2017. 'Estimation of train driver workload: Extracting taskload measures from on-train-data-recorders', in *Communications in Computer and Information Science.* pp.106-119. doi.org/10.1007/978-3-319-61061-0_7

Balfe, N., Sharples, S. and Wilson, J.R., 2015. Impact of automation: Measurement of performance, workload and behaviour in a complex control environment. *Applied ergonomics*, 47, pp.52-64.

Balfe, N., Wilson, J.R., Sharples, S., Clarke, T., 2008. 'Structured observations of automation use', in Bust, P.D. (ed.) *Contemporary Ergonomics*. London:

Taylor & Francis. p.552.
Bergström, J., Van Winsen, R., Henriqson, E., 2015. On the rationale of resilience in the domain of safety: A literature review. *Reliability Engineering & System Safety,* 141, pp.131–141. doi.org/10.1016/j.ress.2015.03.008
Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Zivkovic, V.T., Olmstead, R.E., Tremoulet, P.D., Craven, P.L., 2007. EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks. *Aviation, Space, and Environmental Medicine,* 78(5), pp.B231–B244.
Bluebell Railway, 2021. *The Development of the British Railway Carriage.* http://www.bluebell-railway.co.uk/bluebell/car_fs1.html [Accessed 19.08.2021]
Boas, D.A., Elwell, C.E., Ferrari, M., Taga, G., 2014. Twenty years of functional near-infrared spectroscopy: Introduction for the special issue. *Neuroimage* 85(1), pp.1–5. doi.org/10.1016/j.neuroimage.2013.11.033
Borghini, G., Aricò, P., Di Flumeri, G., Cartocci, G., Colosimo, A., Bonelli, S., Golfetti, A., Imbert, J.P., Granger, G., Benhacene, R., Pozzi, S., Babiloni, F., 2017. EEG-Based Cognitive Control Behaviour Assessment: an Ecological study with Professional Air Traffic Controllers. *Scientific Reports* 7(1), p.547. doi.org/10.1038/s41598-017-00633-7
Borghini, G., Vecchiato, G., Toppi, J., Astolfi, L., Maglione, A., Isabella, R., Caltagirone, C., Kong, W., Wei, D., Zhou, Z., Polidori, L., Vitiello, S., Babiloni, F., 2012. Assessment of mental fatigue during car driving by using high resolution EEG activity and neurophysiologic indices, in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS.* pp.6442–6445. doi.org/10.1109/EMBC.2012.6347469
Boucsein, W., 2012. *Electrodermal activity.* 2nd edition. doi.org/10.1007/978-1-4614-1126-0
Bound, K., 2016. *'Terror & Tension' Psychophysiological Suspense: Defining a framework to measure cinematic suspense in 21 st century horror films,* PhD Thesis, University of Nottingham.
Bradley, S., 2016. *The Railways: Nation, Network and People*. Profile Books
Braithwaite, J.J., Watson, D.G., Jones, R. and Rowe, M., 2013. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology*, 49(1), pp.1017-1034.
Braun, V. and Clarke, V., 2012. 'Thematic analysis', in *APA Handbook of Research Methods in Psychology. Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* pp.57-71. Washington, DC, US: American Psychological Association.
Broekhoven, R.F.G. van, 2016. *Comparison of real-time relative workload measurements in rail signallers.* Masters Thesis. University of Twente.
Broekhoven, R. van, Siegel, A.W., Schraagen, J.M.C., Noordzij, M.L., 2016. Comparison of real-time relative workload measurements in rail signallers, in *Rail Human Factors Proceedings of the 2nd Germany*

Workshop March, 8th and 9th, 2016, Stadthalle Braunschweig. pp.30–40.

Brookhuis, K.A., de Waard, D., 2010. Monitoring drivers' mental workload in driving simulators using physiological measures. *Accidental Analysis & Prevention* 42, pp.898–903. doi.org/10.1016/j.aap.2009.06.001

Brookings, J.B., Wilson, G.F., Swain, C.R., 1996. Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psycho*logy 42, pp.361–377. doi.org/10.1016/0301-0511(95)05167-8

Buksh, A., Sharples, S., Wilson, J.R., Coplestone, A., Morrisroe, G., 2013. 'A comparative cognitive task analysis of the different forms of driving in the UK rail system' in Dashashi, N., Scott, A., Wilson, J.R., Mills, A. (eds.) *Rail Human Factors: Supporting Reliability, Safety and Cost Reduction*. pp.173-182.

Bullough, R. V., Baughman, K., 1995. Changing contexts and expertise in teaching: First-year teacher after seven years. *Teaching and Teacher Education,* 11(5), pp.461-477. doi.org/10.1016/0742-051X(94)00005-Q

Bunce, S.C., Izzetoglu, M., Izzetoglu, K., Onaral B., Pourrezaei, K., 2006. Functional near-infrared spectroscopy. *IEEE Engineering in Medicine and Biology Mag*azine 25(4), pp.54–62. doi.org/10.1109/MEMB.2006.1657788

Cacioppo, J.T., Tassinary, L.G., Berntson, G.G., 2017. *Handbook of psychophysiology*. 4th edition. Cambridge: Cambridge University Press. doi.org/10.1017/9781107415782

Caldwell, J.A., Mallis, M.M., Caldwell, J.L., Paul, M.A., Miller, J.C., Neri, D.F., 2009. Fatigue countermeasures in aviation. *Aviation, Space, and Environmental Medicine*, 80(1), pp.29-59. doi.org/10.3357/ASEM.2435.2009

Cambridge Dictionary, 2021. https://dictionary.cambridge.org/dictionary/english [Accessed 24.03.2021]

Car Sales, 2018. *Fatigue monitoring systems explained*. 5 March 2018. https://www.carsales.com.au/editorial/details/fatigue-monitoring-systems-111391/ [Accessed 10.08.2018]

Cellier, J.M., Eyrolle, H., Mariné, C., 1997. Expertise in dynamic environments. *Ergonomics,* 40(1), pp.28-50. doi.org/10.1080/001401397188350

Charles, R.L. and Nixon, J., 2019. Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics,* 74, pp.221-232. doi.org/10.1016/j.apergo.2018.08.028

Chase, W.G. and Simon, H.A., 1973. Perception in chess. *Cognitive Psychology*, 4 (1), pp.55–81.

Chi, M.T., Glaser, R., Farr, M.J., 1988. *The Nature of Expertise.* Psychology Press. doi.org/10.4324/9781315799681

Collet, C., Salvia, E., Petit-Boulanger, C., 2014. Measuring workload with electrodermal activity during common braking actions. *Ergonomics* 57(6), pp.886–96. doi.org/10.1080/00140139.2014.899627

Colquhoun, H.L., Levac, D., O'Brien, K.K., Straus, S., Tricco, A.C., Perrier, L., Kastner, M., Moher, D., 2014. Scoping reviews: Time for clarity in definition, methods, and reporting. *Journal of Clinical Epidemiology*, 67(12), pp.1291-1294. doi.org/10.1016/j.jclinepi.2014.03.013

Colzato, L.S. and Steenbergen, L., 2017. High vagally mediated resting-state heart rate variability is associated with superior action cascading. *Neuropsychologia,* pp.1–6. doi.org/10.1016/j.neuropsychologia.2017.08.030

Conklin, T., 2012. *Pre-Accident Investigations: An Introduction to Organizational Safety.* Ashgate Publishing Ltd. doi.org/10.1080/00140139.2013.793041

Cooper, M.D., 2020. The emperor has no clothes: A critique of safety-II. *Safety Science*, 152, 105047. doi.org/10.1016/j.ssci.2020.105047

Coveney, C.M., 2014. Managing sleep and wakefulness in a 24-hour world. *Sociology of Health & Illness*, 36 (1), pp.123–136. doi.org/10.1111/1467-9566.12046

Crowley, K. and Balfe, N., 2018. Investigation of Train Driver Physiological Responses, in *Proceedings of the 32nd International BCS HCI Conference.* BCS Learning & Development Ltd., pp.1–9. doi.org/10.14236/ewic/hci2018.5

Darwent, D., Dawson, D., Paterson, J.L., Roach, G.D., Ferguson, S.A., 2015. Managing fatigue: It really is about sleep. *Accident Analysis & Prevention,* 82, pp.20-26. doi.org/10.1016/j.aap.2015.05.009

Davis, F. D., 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13 (3), pp.319–340. doi:10.2307/249008

Davies, D.R. and Parasuraman, R., 1982. *The psychology of vigilance.* Academic Press.

Dawson, D. and McCulloch, K., 2005. Managing fatigue: It's about sleep. *Sleep Medicine Reviews,* 9(5), pp.365-380. doi.org/10.1016/j.smrv.2005.03.002

Dawson, M.E., Schell, A.M., Filion, D.L., 2017. The electrodermal system, in Cacioppo, J.T, Tassinary, L.G., Berntson, G.G. (eds.) *Handbook of Psychophysiology*, Fourth Edition. pp.217–243. Cambridge: Cambridge University Press.

de Greef, T., Lafeber, H., van Oostendorp, H., Lindenberg, J., 2009. Eye Movement as Indicators of Mental Workload to Trigger Adaptive Automation. *HCI* (16), pp.219–228.

de Groot, A., 1965. *Thought and Choice in Chess*, 2nd Edition. The Hague: Mouton Publishers.

Dekker, S., 2014. *Safety Differently: Human Factors for a New Era*. 2nd Edition. London: CRC Press. doi.org/10.1201/b17126

Delamare, L., Lumby, A., Treble, P., Golightly, D., 2016. 'D-MOD Dynamic Modelling of Operator Demand: A new simulator module for the evaluation of signaler's demand', in *2016 IEEE International Conference on Intelligent Rail Transportation (ICIRT)*, pp.362-370. doi.org/10.1109/ICIRT.2016.7588755

Dember, W.N., Warm, J.S., Nelson, W.T., Simons, K.G., Hancock, P.A., Gluckman, J.P., 1993. 'The rate of gain of perceived workload in sustained attention' in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. vol 37. pp.1388–1392. Santa Monica: SAGE Publications.

Derby Signalling, 2021. Signal boxes. http://www.derby-signalling.org.uk/HTML_List/Signalboxes_P.html [Accessed 24.08.2021]

deWaard, 1996. *The Measurement of Drivers' Mental Workload.* Traffic Research Centre, University of Groningen.

Dirican, A.C. and Gokturk, M., 2011. Psychophysiological measures of human cognitive states applied in Human Computer Interaction. *Procedia Computer Science* 3, pp.1361–1367. doi.org/10.1016/j.procs.2011.01.016

Dorrian, J., Baulk, S.D., Dawson, D., 2011. Work hours, workload, sleep and fatigue in Australian Rail Industry employees. *Applied Ergonomics,* 42(2), pp.202-209. doi.org/10.1016/j.apergo.2010.06.009

Dreyfus, H.L. and Dreyfus, S., 1986. *Mind over Machine: the power of human intuition and expertise in the age of the computer.* Oxford: Basil Blackwell.

Dunn, N. and Williamson, A., 2012. Driving monotonous routes in a train simulator: the effect of task demand on driving performance and subjective experience. *Ergonomics*, 55(9), pp.997–1008. doi.org/10.1080/00140139.2012.691994

Edwards, T., Sharpies, S., Kirwan, B., Wilson, J., 2016. 'Identifying markers of performance decline in air traffic controllers', in Di Bucchianico,G., Vallicelli, A., Stanton, N.A., Landry, S.J., (eds.) *Human Factors in Transportation: Social and Technological Evolution Across Maritime, Road, Rail, and Aviation Domains.* Boca Raton: CRC Press. doi.org/10.1201/9781315370460

Eggemeier, F.T. and Wilson, G.F., 1991. 'Performance-based and subjective assessment of workload in multi-task', in Damos, D. (ed) *Multiple task performance*. London: CRC Press. p.217.

Ericsson, K. and Smith, J., 1991. 'Prospects and limits of the empirical study of expertise: an introduction', in Ericsson, K.A. and Smith, J. (eds.) *Toward a general theory of expertise: Prospects and limits*, pp.01-38. Cambridge: Cambridge University Press.

Ericsson, K. A. and Smith, J., 1991. Prospects and limits of the empirical study of expertise: An introduction. In K. A. Ericsson & J. Smith (eds.), *Toward a general theory of expertise: Prospects and limits*, pp. 1–38. Cambridge: Cambridge University Press.

ERTMS, 2021. *ERTMS provides the European Union with a unique opportunity to create a seamless railway system*. https://www.ertms.net/ [Accessed 25.08.2021]

Farrington-Darby, T., Wilson, J.R., 2006. The nature of expertise: A review. *Applied Ergonomics,* 37(1), pp.17-32. doi.org/10.1016/j.apergo.2005.09.001

Filtness, A.J. and Naweed, A., 2017. Causes, consequences and countermeasures to driver fatigue in the rail industry: The train driver perspective. *Applied Ergonomics,* 60, pp.12–21. doi.org/10.1016/j.apergo.2016.10.009

Flin, R., O'Connor, P., Crichton, M., 2017. *Safety at the Sharp End, Safety at the Sharp End.* London: CRC Press. doi.org/10.1201/9781315607467

Forte, G., Favieri, F., Casagrande, M., 2019. Heart rate variability and cognitive function: A systematic review. *Frontiers in Neuroscience*, 13, p.710.

doi.org/10.3389/fnins.2019.00710

Fortune Business Insights, 2021. *Fitness Tracker Market Size, Share & COVID-19 Impact Analysis.* https://www.fortunebusinessinsights.com/fitness-tracker-market-103358 [Accessed 22/10/21]

Fowler, A., Golightly, D., Sharples, S., Harvey, C., Wilson, M.L., Gibson, H., 2019. 'Human performance in rail: current assessment and future opportunities', in Charles, R., and Golightly, D. (eds.) *Contemporary EHF 2019*. ISBN 978-1-9996527-1-5

Fox A, 1974. *Beyond contract: Work, power and trust relations.* London: Faber and Faber.

Fujiyama T, 2018. *Comparing railway systems in the UK and Japan from the view of their punctuality*. https://dajf.org.uk/wp-content/uploads/Daiwa-Seminar.pdf [Accessed 04/01/2023]

Fuster, Joaquin, M., 2001. The Prefrontal Cortex — An Update: Time is of the Essence. *Neuron*, 30(2), pp.319–333.

Gao, Q., Wang, Y., Song, F., Li, Z., Dong, X., 2013. Mental workload measurement for emergency operating procedures in digital nuclear power plants. *Ergonomics*, 56(7), pp.1070–1085. doi.org/10.1080/00140139.2013.790483

Gelfand, M.J., Jackson, J.C., Pan, X., Nau, D., Pieper, D., Denison, E., Dagher, M., Van Lange, P.A.M., Chiu, C.Y., Wang, M., 2021. The relationship between cultural tightness–looseness and COVID-19 cases and deaths: a global analysis. *The Lancet Planetary Health*, 5(3), pp.e135–e144. doi.org/10.1016/S2542-5196(20)30301-6

Gertler, J. and Viale, A., 2007. *Work schedules and sleep patterns of railroad dispatchers* (No. DOT/FRA/ORD-07/11). www.railroads.dot.gov/elibrary/work-schedules-and-sleep-patterns-railroad-dispatchers [Accessed 27.10.2022]

Gillis, I., 2016. 'Cognitive workload of train drivers', in Wilson, J.R., Norris, B., Clarke, T., Mills, A., (eds.) *People and Rail Systems: Human Factors at the Heart of the Railway*. Ashgate, pp.91–101. doi.org/10.4324/9781315247182

Glaser, R., 1987. 'Thoughts on Expertise' in Schooler, C. and Shaie, W. (eds.) *Cognitive Functioning and Social Structure over the Life Course*.

Glaus, A., Crow, R., Hammond, S., 1996. A qualitative study to explore the concept of fatigue/tiredness in cancer patients and in healthy individuals. Supportive Care in Cancer 4, pp.82–96.

Golightly, D., Ryan, B., Dadashi, N., Pickup, L., Wilson, J.R., 2013. Use of scenarios and function analyses to understand the impact of situation awareness on safe and effective work on rail tracks. *Safety Science*, 56, pp.52-62. doi.org/10.1016/j.ssci.2012.08.007

Golightly, D. and Young, M.S., 2022. Local knowledge in rail signalling and balancing trade-offs. *Applied Ergonomics,* 102. p.103714. doi.org/10.1016/j.apergo.2022.103714

Gribel, L., Regier, S., Stengel, I., 2016. 'Acceptance factors of wearable computing: An empirical investigation', in *Proceedings of the 11th International Network Conference (INC)*, pp.67-72.

Grover, M.D., 1983. Pragmatic Knowledge Acquisition Methodology. In *INCAI*, vol.83, pp. 436-438.

Hack, C., 2021. 'Protecting track workers from trains: Understanding signallers' challenges in blocking the line between trains', *Seventh International Human Factors Rail Conference. Live virtual conference*, 23-25 June 2021. https://www.ukrrin.org.uk/events-and-activity/the-seventh-international-rail-human-factors-rail-conference/ [Accessed 09.02.2023]

Hamilton, W.I. and Clarke, T., 2005. Driver performance modelling and its practical application to railway safety. *Applied Ergonomics,* 36, pp.661–670. doi.org/10.1016/J.APERGO.2005.07.005

Hankins, T.C. and Wilson, G.F., 1998. A comparison of heart rate, eye activity, EEC and subjective measures of pilot mental workload during flight. *Aviation, Space, and Environmental Medicine*, 69, pp.360–367. doi.org/10.1519/JSC.0b013e3182542674

Harmat, L., de Manzano, T., Theorell, T., Högman, L., Fischer, H., Ullén, F., 2015. Physiological correlates of the flow experience during computer game playing. *International Journal of Psychophysiology*, 97, pp.1–7.

Hart, S.G. and Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, vol. 52, pp.139-184. doi.org/10.1016/S0166-4115(08)62386-9

Healey, J.A. and Picard, R.W., 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Tranactions on Intelligent Transportation Systems*, 6(2), pp.156–166. doi.org/10.1109/TITS.2005.848368

Hendy, K.C., 1997. An information processing approach to workload and situation awareness. In Jensen, R.S., and Rakovan, L.A. (eds.) *Proceedings of the 9th International Symposium on Aviation Psychology*, pp.1423-1428.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., Schultz, T., 2013. Mental workload during n-back task-quantified in the prefrontal cortex using fNIRS. *Frontiers in Human Neuroscience*, 7, p.935. doi.org/10.3389/fnhum.2013.00935

Hignett, S., 2001. *Using Qualitative Methodology in Ergonomics: theoretical background and practical examples*, PhD thesis, University of Nottingham.

Hoffman, R.R., 2014. *The Psychology of Expertise: Cognitive research and empirical AI*. London: Psychology Press. doi.org/10.4324/9781315806105

Hoffman, R.R., 1998. 'How Can Expertise be Defined? Implications of Research from Cognitive Psychology' in Williams, R., Faulkner, W., Fleck, J. (eds.) *Exploring Expertise*. pp. 81-100. doi.org/10.1007/978-1-349-13693-3_4

Hoffman, R.R., Shadbolt, N.R., Burton, A.M., Klein, G., 1995. Eliciting knowledge from experts: A methodological analysis. *Organanizational behaviour and human decision processes* 62(2), pp.129-158. doi.org/10.1006/obhd.1995.1039

Hollnagel, E., 2017. *Safety-II in practice: Developing the resilience potentials,* *Safety-II in Practice: Developing the Resilience Potentials.* Abingdon, UK: Routledge. doi.org/10.4324/9781315201023

Hollnagel, E., 2014. *Safety-I and Safety-II, the past and future of safety management*. London: CRC Press.

Hollnagel, E., Leonhardt, J., Licu, T., Shorrock, S., 2013. *From Safety-I to Safety-II: A White Paper*. European Organisation for the Safety of Air Navigation. Brussels.

Holtermann, A., Mathiassen, S.E. and Straker, L., 2019. Promoting health and physical capacity during productive work: the Goldilocks Principle. *Scandinavian journal of work, environment & health*, 45(1), pp.90-97.

Hugdahl, K., 1995. *Psychophysiology: The mind-body perspective*. Harvard University Press.

Hurt, H.T., Joseph, K., Cook, C.D., 1977. Scales for the Measurement of Innovativeness. *Human Communication Research*, 4(1), pp.58–65. doi.org/10.1111/j.1468-2958.1977.tb00597.x

Jacobs, J. V., Hettinger, L.J., Huang, Y.H., Jeffries, S., Lesch, M.F., Simmons, L.A., Verma, S.K., Willetts, J.L., 2019. Employee acceptance of wearable technology in the workplace. *Applied Ergonomics*, 78. doi.org/10.1016/j.apergo.2019.03.003

Jarosz, I., Lo, J., Lijs, J., 2021. Investigating the role of Non-Technical Skills in Railway Traffic Operations through Expert Knowledge Elicitation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 65, No. 1, pp.262-266. Los Angeles: Sage CA. doi.org/10.1177/1071181321651181

Johnsen, B.H., Thayer, J.F., Laberg, J.C., Wormnes, B., Raadal, M., Skaret, E., Kvale, G., Berg, E., 2003. Attentional and physiological characteristics of patients with dental anxiety. *Journal of Anxiety Disorders*. 17, pp.75–87. doi.org/10.1016/S0887-6185(02)00178-0

Jones, C.B., Dorrian, J., Rajaratnam, S.M. and Dawson, D., 2005. Working hours regulations and fatigue in transportation: A comparative analysis. *Safety science* 43(4), pp.225-252.

Jordan, C. S., Farmer, E. W. and Belyavin, A. J., 1995. The DRA Workload Scale: A Validated Workload Assessment Technique. *Proceedings of 8th International Symposium on Aviation Psychology*. Vol 2, pp.1013-1018.

Kahneman, D., 1973. ‘Toward a theory of mental effort’, in D. Kahneman (ed). *Attention and effort*. Prentice-Hall. pp.13-27.

Kamzanova, A.T., Kustubayeva, A.M., Matthews, G., 2014. Use of EEG Workload Indices for Diagnostic Monitoring of Vigilance Decrement. *Human Factors*, 56, pp.1136–1149. doi.org/10.1177/0018720814526617

Kilpatrick, D.G., 1972. Differential Responsiveness of Two Electrodermal Indices to Psychological Stress and Performance of a Complex Cognitive Task. *Psychophysiology*, 9(2): 218-226. doi.org/10.1111/j.1469-8986.1972.tb00756.x

Kim, J-Y., Cho, Y-J., and Min, S-N., 2015. ‘Psychophysiological measurement of physical and cognitive work’, in Wilson, J., Sharples, S. (eds.) *Evaluation of Human Work.* 4th Edition. London: CRC Press, pp.583-607.

Kim, J. and Courtney, J.F., 1988. A survey of knowledge acquisition techniques and their relevance to managerial problem domains. *Decision Support Systems*, 4(3), pp.269-284. doi.org/10.1016/0167-9236(88)90016-4

Kim, S.H., Lim, K.R., Seo, J.H., Ryu, D.R., Lee, B.K., Cho, B.R. and Chun, K.J., 2022. Higher heart rate variability as a predictor of atrial fibrillation in patients with hypertension. *Scientific Reports,* 12(1), p.3702.

Klein, G.A., 1993. A recognition-primed decision (RPD) model of rapid decision making. *Decision making in action: Models and methods*, 5(4), pp.138-147.

Klein, M. and Malzahn, D., 1991. 'Effects of physical ability on working memory requirements, keystroke rate, and subjective workload of computer input devices', in *Proceedings of the Human Factors Society Annual* Meeting (Vol. 35, No.5, pp.261-265). Los Angeles, CA: Sage CA. doi.org/10.1177/154193129103500504

Knight, J. and Salvendy, G., 1981. Effects of task feedback and stringency of external pacing on mental load and work performance. *Ergonomics*, 24(10), pp.757-764. doi.org/10.1080/00140138108924897

Kolodner, J.L., 1991. Improving human decision making through case-based decision aiding. *AI Magazine*, 12(2), pp.52-68. doi:10.1609/aimag.v12i2.895

Krehl, C. and Balfe, N., 2014. Cognitive workload analysis in rail signalling environments. *Cognition, Technology & Work* 16, pp.359–371. doi.org/10.1007/s10111-013-0266-7

Kyriakidis, M., Majumdar, A., Ochieng, W.Y., 2015. Data based framework to identify the most significant performance shaping factors in railway operations. *Safety Science,* 78, pp.60-76. doi.org/10.1016/j.ssci.2015.04.010

Lacey, J.I., Kagan, J., Lacey, B.C., Moss, H.A., 1963. 'The visceral level: Situational determinants and bevaioural correlates of autonomic response patterns', in Darwin, C., *Expression of the Emotions in Man*. New York: International Universities Press, pp.161-196.

Large, D.R., Golightly, D., Taylor, E.L., 2014. 'The effect of driver advisory systems on train driver workload and performance' in Sharples, S. and Shorrock, S (eds.) *Contemporary Ergonomics and Human Factors 2014.* Abingdon, UK: Routledge. doi.org/10.1201/b16742

Larichia, F., 2022a. *Fitbit – statistics & facts*. https://www.statista.com/topics/2595/fitbit/#topicOverview [Accessed 04/02/2023]

Larichia, F., 2022b. *Wearables – statistics & facts*. https://www.statista.com/topics/1556/wearable-technology/#topicOverview [Accessed 04/01/2023]

Larue, G.S., Rakotonirainy, A., Haworth, N.L., 2016. A simulator evaluation of effects of assistive technologies on driver cognitive load at railway-level crossings. *Journal of Transportation Safety and Security,* 8(1), pp.56-69. doi.org/10.1080/19439962.2015.1055413

Lehrer, P, Karavidas, M., Lu, S., Vaschillo, E., Vaschillo, B., Cheng, A., 2010. Cardiac data increase association between self-report and both expert ratings of task load and task performance in flight simulator tasks: An exploratory study. *International Journal of Psychophysiology,* 76, pp.80–87. doi.org/10.1016/j.ijpsycho.2010.02.006

Lehto, M.R. and Buck, J.R., 1988. Status reporting of “eyes on” and peripheral displays during tracking. *Perceptual and Motor Skills,* 67(3), pp.719-733. doi.org/10.2466/pms.1988.67.3.719

Lenat, D.B. and Feigenbaum, E.A., 1988. ‘On the thresholds of knowledge’ in *Proceedings of the International Workshop on Artificial Intelligence for Industrial Applications.* 25-27 May 1988, Hitachi City, Japan. IEEE, pp. 291-300. doi.org/10.1109/aiia.1988.13308

Lingard, H., Wakefield, R., Blismas, N., 2013. “ If you cannot measure it , you cannot improve it ”: Measuring health and safety performance in the construction industry, in *19th Triennial CIB World Building Congress.* Brisbane, Australia.

Lowe, E. and Pickup, L., 2008. ‘Network rail signaller’s workload toolkit’, in *Contemporary Ergonomics 2008*. p.558. doi.org/10.1201/9780203883259

Mackworth, N.H., 1948. The Breakdown of Vigilance during Prolonged Visual Search. *Quarterly journal of experimental psychology,* 1(1), pp.6–21. doi.org/10.1080/17470214808416738

Madigan, R., Golightly, D., Madders, R., 2016. Application of Human Factors Analysis and Classification System (HFACS) to UK rail safety of the line incidents. *Accident Analysis & Prevention,* 97, pp.122-131. doi.org/10.1016/j.aap.2016.08.023

Madigan, R., Golightly, D., Madders, R., 2015. ‘Rail industry requirements around non-technical skills’, in *Contemporary Ergonomics and Human Factors.* pp.474-481. doi.org/10.1201/b18293-70

Maior, H.A., Wilson, M.L., Locke, C. and Swann, D., 2018. Brain activity and mental workload associated with artistic practice. *CHI 2018 Artistic BCI Workshop*. *ACM CHI Conference on Human Factors in Computing Systems (CHI 2018).* 21-26 April 2018. Montreal, Canada.

Marinescu, A.C., Sharples, S., Ritchie, A.C., Sánchez López, T., McDowell, M. and Morvan, H.P., 2018. Physiological parameter response to variation of mental workload. *Human Factors,* 60(1), pp.31–56.

Marinescu, A., Sharples, S., Ritchie, A.C., Sánchez López, T., McDowell, M., Morvan, H., 2016. Exploring the Relationship between Mental Workload, Variation in Performance and Physiological Parameters. *IFAC-PapersOnLine,* 49(19), pp.591–596. doi.org/10.1016/j.ifacol.2016.10.618

Martindale, V.E., 2012. Breathalyzer for fatigue: The “Fatigualyzer.” *Aviation, Space, and Environmental Medicine.* 83(1), pp.70-71. doi.org/10.3357/ASEM.3210.2012

Matthews, G., Reinerman-Jones, L.E., Barber, D.J., Abich, J., 2014. The Psychometrics of Mental Workload: Multiple Measures Are Sensitive but Divergent. *Human Factors,* 57(1), pp.125–143. doi.org/10.1177/0018720814539505

May, J.F. and Baldwin, C.L., 2009. Driver fatigue: The importance of identifying causal factors of fatigue when considering detection and countermeasure technologies. *Transportation research part F: traffic* *psychology and behaviour,* 12(3), pp.218-224. doi.org/10.1016/j.trf.2008.11.005

McDonald, A., 2018. Maximising human performance for future systems. Keynote presentation to *13th International Symposium of AAvPA and PACDEFF 2018*. 8 November 2018.

Mehler, B., Reimer, B., Coughlin, J.F., 2012. Sensitivity of Physiological Measures for Detecting Systematic Variations in Cognitive Demand From a Working Memory Task. *Human Factors* 54, pp.396–412. doi.org/10.1177/0018720812442086

Meshkati, N., Hancock, P. a., Rahimi, M., Dawes, S.M., 1995. Techniques in mental workload assessment. *Eval. Hum. Work A Pract. Ergon. Methodol.*

Mesulam, M. M, 1990. Large-scale neurocognitive networks and distributed processing for attention, language, and memory. Ann. Neurol. 28, pp.597–613. doi.org/10.1002/ana.410280502

Milton, N., 2003. Personal Knowledge Techniques. University of Nottingham, Nottingham.

Moray, N., 2008. The good, the bad, and the future: On the archaeology of ergonomics. *Human Factors,* 50(3), pp.411-417.

Mühl, C., Jeunet, C., Lotte, F., 2014. EEG-based workload estimation across affective contexts. *Frontiers in Neuroscience.* doi.org/10.3389/fnins.2014.00114

Naweed, A. and Rose, J., 2018. Assessing technology acceptance for skills development and real-world decision-making in the context of train driving. Transp. Res. Part F Traffic Psychol. Behav. 52. doi.org/10.1016/j.trf.2017.11.003

Nagano, Y., 2002. Change in baroreflex sensitivity during mirror drawing: effect of task difficulty. *Japanese Journal of Physiological Psychology and Psychophysiology* 20, pp.233–239.

Nardi, B.A. and Reilly, B., 1997. Interactive ethnography: digital photography at Lincoln High School. In *CHI'97 Extended Abstracts on Human Factors in Computing Systems*, pp.18-19.

Naweed, A., 2014. Investigations into the skills of modern and traditional train driving. *Applied Ergonomics,* 45(3), pp.462-470.

Naweed, A. and Aitken, J., 2014. 'Drive a mile in my seat: Signal design from a systems perspective'. *Technical Meeting of the Institution of Railway Signal Engineers*, July 2014, Newcastle.

Nayak, S., Tripathy, S., Dash, A., 2018. Role of non technical skill in human factor engineering: a crucial safety issue in Indian Railway. *International Journal of System Assurance Engineering Management* 9, pp.1120-1136. doi.org/10.1007/s13198-018-0715-z

Network Rail, 2019. *Infrastructure Insights: Signalling*. April 20, 2019 https://www.networkrail.co.uk/stories/infrastructure-insights-signalling/ [Accessed 19.8.2021]

Network Rail, 2023. *Delay Attribution Board.* https://www.networkrail.co.uk/industry-and-commercial/information-for-operators/network-code/delay-attribution-board/ [Accessed 23.02.23]

Nguyen, T., Ahn, S., Jang, H., Jun, S.C., Kim, J.G., 2017. Utilization of a combined EEG/NIRS system to predict driver drowsiness. Sci. Rep. 7. doi.org/10.1038/srep43933

Nickel, P. and Nachreiner, F., 2003. Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload. *Human Factors* 45, pp.575–590. doi.org/10.1518/hfes.45.4.575.27094

Oken, B.S., Salinsky, M.C., Elsas, S.M., 2006. Vigilance, alertness, or sustained attention: physiological basis and measurement. doi.org/10.1016/j.clinph.2006.01.017

ONRSR, 2021. *Managing fatigue.* Office of the National Rail Safety Regulator, Australia.

ORR, 2012. *Managing Rail Staff Fatigue*. January 2012. file:///Users/psxacf/Downloads/managing_rail_fatigue%20(1).pdf [Accessed 10.12.2022]

ORR, 2013. *Health and safety report 2013*. 22 December 2013. https://www.orr.gov.uk/media/20088 [Accessed 10.12.2022]

ORR, 2017. *Fatigue: Key Performance Indicators*. 11 May 2017. https://www.orr.gov.uk/media/10939 [Accessed 10.12.2022]

ORR, 2018. *Passenger Rail Usage 2018-2019 Q1 Statistical Release.* http://orr.gov.uk/__data/assets/pdf_file/0003/39342/passenger-rail-usage-2018-19-q1.pdf [Accessed 23.10.18]

ORR, 2020. Rail Infrastructure and Assets https://dataportal.orr.gov.uk/media/1842/rail-infrastructure-assets-2019-20.pdf [Accessed 19.08.2021]

ORR, 2021. ERMTS https://www.orr.gov.uk/guidance-compliance/rail/health-safety/infrastructure/train-protection/ertms [Accessed 25.08.2021]

ORR, 2022. Passenger rail performance 1 January to 31 March 2022. https://dataportal.orr.gov.uk/media/2061/passenger-performance-jan-mar-2022.pdf [Accessed 04/01/2023]

Palmqvist, C.W., Olsson, N.O. and Hiselius, L.W., 2017. Some influencing factors for passenger train punctuality in Sweden. *International Journal of Prognostics and Health Management* 8(3).

Parasuraman, A. and Colby, C.L., 2015. An Updated and Streamlined Technology Readiness Index: TRI 2.0. *Journal of Service Research.* 18(1). doi.org/10.1177/1094670514539730

Parasuraman, R., Riley, V., 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors,* 39(2). doi.org/10.1518/001872097778543886

Parasuraman, R., Warm, J.S., See, Judi, E., 1998. 'Brain Systems of Vigilance', in Parasuraman, R. (ed.), *The Attentive Brain.* MIT Press, Cambridge, MA, pp.221–256.

PENNA, 2018. *Network Rail Signallers Job Analysis*. December 2018

Phillips, R.O., and Sagberg, F., 2010. *Managing driver fatigue in occupational settings*. TØI report 1081/2010. Oslo: Institute of Transport Economics.

Pickup, L., 2006. *Understanding and assessment of mental workload in railway signalling,* PhD thesis, University of Nottingham.

Pickup, L. and Wilson, J. R., 2007. Mental Workload Assessment and the

Development of the Operational Demand Evaluation Checklist (ODEC) for Signallers. In Wilson J. R. (ed.) *People and Rail Systems: Human Factors at the Heart of the Railway*. London: CRC Press.

Pickup, L., Wilson, J., Lowe, E., 2010. The Operational Demand Evaluation Checklist (ODEC) of workload for railway signalling. *Applied Ergonomics*, 41, pp.393–402. doi.org/10.1016/j.apergo.2009.09.003

Pickup, L., Wilson, J.R., Norris, B.J., Mitchell, L., Morrisroe, G., 2005b. The Integrated Workload Scale (IWS): A new self-report tool to assess railway signaller workload. *Applied Ergonomics,* 36, pp.681–693. doi.org/10.1016/j.apergo.2005.05.004

Pickup, L., Wilson, J.R., Sharples, S., Norris, B., Clarke, T. and Young, M.S., 2005a. Fundamental examination of mental workload in the rail industry. *Theoretical issues in ergonomics science* 6(6), pp.463-482.

Pigeau, R.A., Angus, R.G., O'Neill, P., Mack, I., 1995. Vigilance latencies to aircraft detection among NORAD surveillance operators. *Human Factors,* 37. doi.org/10.1518/001872095779049291

Practice4Me 2021 https://www.practice4me.co.uk/network-rail-assessment/ [Accessed 16.08.21]

RAIB, 2020. Class investigation into human performance in signalling operations. Report 03/2020.

Ramírez, E., Ortega, A.R., Reyes Del Paso, G.A., 2015. Anxiety, attention, and decision making: The moderating role of heart rate variability. *Int. J. Psychophysiol.* 98, pp.490–496. doi.org/10.1016/j.ijpsycho.2015.10.007

Rasmussen, J., 1997. Risk management in a dynamic society: A modelling problem. *Saf. Sci.* 27. doi.org/10.1016/S0925-7535(97)00052-0

Rasmussen, J., 1983. Skills, Rules, and Knowledge; Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models. *IEEE Transactions on Systems, Man, and Cybernetics.* SMC-13, pp.257–266. doi.org/10.1109/TSMC.1983.6313160

Rasmussen, J., Brehmer, B., Leplat, J., 1991. *Distributed Decision Making: Cognitive Models for Cooperative Work*. Chichester.

Rasmussen, J., Jensen, A., 1974. Mental procedures in real-life tasks: A case study of electronic trouble shooting. *Ergonomics* 17(3). doi.org/10.1080/00140137408931355

Reason, J., 1990. Human Error. Cambridge University Press.

Reichardt, C.S., Rallis, S.F., 1994. Qualitative and quantitative inquiries are not incompatible: A call for a new partnership. *New Directions for Program Evaluation*, 61, pp.85-91. doi.org/10.1002/ev.1670

Reid, G.B. and Nygren, T.E., 1988. The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. *Advance in Psychol*ogy, 52, pp.185-218. doi.org/10.1016/S0166-4115(08)62387-0

Rendon-Velez, E., Van Leeuwen, P.M., Happee, R., Horváth, I., van der Vegte, W.F. and de Winter, J.C., 2016. The effects of time pressure on driver performance and physiological activity: A driving simulator study. *Transportation research part F: traffic psychology and behaviour* 41, pp.150-169.

Richter, P., Wagner, T., Heger, R., Weise, G., 1998. Psychophysiological analysis of mental load during driving on rural roads-a quasi-experimental field study. *Ergonomics*, 41(5), pp.593–609. doi.org/10.1080/001401398186775

Riganello, F., Garbarino, S., Sannita, W.G., 2012. Heart rate variability, homeostasis, and brain Function: A tutorial and review of application. *Journal of Psychophysiology*, 26, pp.178–203. doi.org/10.1027/0269-8803/a000080

RMT, 2020. *Safety related incidents and signaller workload increase caused by Automatic Route Setting (ARS)*. https://www.rmt.org.uk/about/health-and-safety/health-and-safety-circulars/safety-related-incidents-and-signaller-workload-increase270120/ [Accessed 15.08.2021]

Roach, G.D., Reid, K.J., Dawson, D., 2003. The amount of sleep obtained by locomotive engineers: effects of break duration and time of break onset. *Occupational and Environmental Medicine*, 60(12), pp.e16-e17. doi.org/10.1136/oem.60.12.e17

Robson, C. 1993. *Real World Research*. Oxford: Blackwell Publishers.

Robson, C., McCartan, K., 2015. *Real World Research*, 4th ed. Wiley.

Rogers, E.M., 1962. *Diffusion of Innovations*. Glencoe: Free Press.

Rose, J. and Bearman, C., 2013. 'Resistance to technology', in *Evaluation of Rail Technology: A Practical Human Factors Guide*. Abingdon, UK: Routledge. doi.org/10.1201/9781315581217-3

Roth, E.M., and Patterson, E.S., 2005. 'Using Observational Study as a Tool for Discovery: Uncovering cognitive and collaborative demands and adaptive strategies' in H. Montgomery, R. Lipshitz, & B. Brehmer (eds.) *How professionals make decisions*. Lawrence Erlbaum Associates Publishers. pp.379–393.

Rouse, W.B., Edwards, S.L. and Hammer, J.M., 1993. Modelling the dynamics of mental workload and human performance in complex systems. *IEEE transactions on systems, man, and cybernetics* 23(6), pp.1662-1671.

RSSB, 2005a. *Tools for train driver mental workload assessment: final report – part one*. T147.

RSSB, 2005b. *Tools for train driver mental workload assessment: final report – part two*. T147.

RSSB, 2012. *Non-Technical Skills.* http://www.rssb.co.uk/Library/improving-industry-performance/2012-leaflet-non-technical-skills.pdf [Accessed 18/09/2022]

RSSB, 2017. *Rail Technical Strategy Capability Delivery Plan.*

RSSB, 2018. *Risk-based Training Needs Analysis (RBTNA).* www.rssb.co.uk [Accessed 30.01.21]

RSSB, 2016. A Good Practice Guide to Integrating Non-Technical Skills into Rail Safety Critical Roles. https://www.rssb.co.uk/Library/improving-industry-performance/2016-07-non-technical-skills-integration-good-practice-guide.pdf [Accessed 17.09.18]

Ryan, B., Golightly, D., Pickup, L., Reinartz, S., Atkinson, S., Dadashi, N., 2021. Human functions in safety-developing a framework of goals, human functions and safety relevant activities for railway socio-technical

systems. *Safety Science,* 140. p.105279. doi.org/10.1016/j.ssci.2021.105279
Saldaña, J., 2016. *The coding manual for qualitative researchers.* Sage.
Salter, W.J., 1988. Human Factors in Knowledge Acquisition, in Helander, M. (ed.) *Handbook of Human-Computer Interaction*. pp.957-968. doi.org/10.1016/b978-0-444-70536-5.50050-6
Sanders, J.P., Loveday, A., Pearson, N., Edwardson, C., Yates, T., Biddle, S.J. and Esliger, D.W., 2016. Devices for self-monitoring sedentary time or physical activity: a scoping review. *Journal of Medical Internet Research,* 18(5), p.e90.
Sapolsky, R.M., 1994, August. Individual differences and the stress response. *Seminars in Neuroscience*, Vol. 6, No. 4. pp. 261-269. Academic Press.
Sawin, D.A. and Scerbo, M.W., 1993. 'Vigilance: where has all the workload gone?', in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 37, No. 19, pp.1383-1387). Sage CA: Los Angeles, CA: SAGE Publications. doi.org/10.1518/107118193784162308
Schellekens, J.M., Sijtsma, G.J., Vegter, E., Meijman, T.F., 2000. Immediate and delayed after-effects of long lasting mentally demanding work. *Biological Psychology.* 53(1), pp.37–56.
Schön, D.A., 1983. *The Reflective Practitioner: How Professionals Think in Action.* New York: Basic Books.
Seshadri, D.R., Li, R.T., Voos, J.E., Rowbottom, J.R., Alfes, C.M., Zorman, C.A. and Drummond, C.K., 2019. Wearable sensors for monitoring the physiological and biochemical profile of the athlete. *NPJ digital medicine,* 2(1), p.72.
Selkurt, E.E., 1976. *Physiology.* 4th edition. Little Brown & Company.
Shadbolt, N.R. and Smart, P.R., 2015. 'Knowledge elicitation', in Wilson, J., Sharples, S. (eds.), *Evaluation of Human Work*, 4th Edition. London: CRC Press. pp.163–200.
Shanteau, J., 1992. 'The psychology of experts: An alternative view', in Wright, G., Bolger, F. (eds.) *Expertise and decision support*. Plenum Press, New York, pp.11–23.
Sharples, S., 2019. 'Workload II: A future paradigm for analysis and measurement', in *Advances in Intelligent Systems and Computing*. Springer. pp.489–498. doi.org/10.1007/978-3-319-96071-5_52
Sharples, S. and Megaw, T., 2015. 'The definition and measurement of mental workload', in *Evaluation of Human Work*, 4th Edition. London: CRC Press. pp.515-548. doi.org/10.1201/b18362
Sharples, S., Millen, L., Golightly, D., Balfe, N., 2011. 'The impact of automation in rail signalling operations', in *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit.* Sage Publications. doi.org/10.1177/09544097JRRT380
Sherwood, L., 2013. *Introduction to Human Physiology*, 8th ed. Andover: Brooks/Cole.
Slade, S., 1991. Case-Based Reasoning: A Research Paradigm. *AI Magazine*. 12(1), pp.42-42.
Sokolov, Y.. N., 1963. *Perception and the conditioned reflex.* London:

Pergamon Press.

Song, Y.S., Baek, J.H., Hwang, D.S., Lee, J.W., Lee, Y.J., Park, H.J., Choi, J.H., Yang, H.K., 2014. Physiological status assessment of locomotive engineer during train operation. *Journal of Electrical Engingeering and Technology*, 9(1), pp.324-333. doi.org/10.5370/JEET.2014.9.1.324

Soon, S., Svavarsdottir, H., Downey, C. and Jayne, D.G., 2020. Wearable devices for remote vital signs monitoring in the outpatient setting: an overview of the field. *BMJ Innovations,* 6(2).

St Clair Gibson, A., Swart, J., Tucker, R., 2018. The interaction of psychological and physiological homeostatic drives and role of general control principles in the regulation of physiological systems, exercise and the fatigue process – The Integrative Governor theory. *European Journal of Sport Science,* 18(1), pp.25-36. doi.org/10.1080/17461391.2017.1321688

Strauss, A. and Corbin, J.M., 1990. *Basics of qualitative research: Grounded theory procedures and techniques.* Sage Publications, Inc.

Sussman, D. and Coplen, M., 2000. Fatigue and alertness in the United States railroad industry part I: the nature of the problem. *Transportation research part F: traffic psychology and behaviour,* 3(4), pp.211-220.

Swart, J., Lindsay, T.R., Lambert, M.I., Brown, J.C. and Noakes, T.D., 2012. Perceptual cues in the regulation of exercise performance–physical sensations of exercise and awareness of effort interact as separate cues. *British Journal of Sports Medicine,* 46(1). pp.42-48.

Tang, X., Zhou, P., Wang, P., 2016. 'Real-time image-based driver fatigue detection and monitoring system for monitoring driver vigilance', in *35th Chinese Control Conference, (CCC)*. pp. 4188–4193. IEEE. doi.org/10.1109/ChiCC.2016.7554007

Tattersall, A.J. and Foord, P.S., 1996. An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics,* 39(5), pp.740-748.

Thomas-Friedrich, B., Schneider, P., Herholz, H. and Grippenkoven, J.D., 2017. 'Measuring Rail Signaller Workload in a highly realistic simulated environment', in *Proceedings of the Sixth International Human Factors Rail Conference*, pp.286-295.

Thompson, K., 2017. Qualitative research rules: Using qualitative and ethnographic methods to access the human dimensions of technology. In *Evaluation of Rail Technology*. pp.75-110. London: CRC Press.

Thorne, E. and Rawlinson, L., 2021. A workload assessment methodology to identify potential for task conflict for control of multiple CCTV level crossings. *Seventh International Human Factors Rail Conference*. 23-25 June 2021. Rail Safety and Standards Board. Live virtual conference. https://www.ukrrin.org.uk/events-and-activity/the-seventh-international-rail-human-factors-rail-conference/ [Accessed 09.02.2023]

Thoroman, B., Goode, N., Salmon P., and Wooley, M., 2019. What went right? An analysis of the protective factors in aviation near misses. *Ergonomics,* 62(2), pp.192-203.

Tortora, G., and Derrickson, B., 2007. *Introduction to the Human Body.* John Wiley & Sons.

Train Driver.org, 2021. The Train Driver Assessment & Interviews. https://www.traindriver.org/assessments-interviews.html# [Accessed 16.08.2021]

Transport Canada, 2022. *Let's talk fatigue in the Rail industry – Consultation on proposed fatigue management system regulations*. https://tc.canada.ca/en/corporate-services/consultations/let-s-talk-fatigue-rail-industry-consultation-proposed-fatigue-management-system-regulations [Accessed 10.12.2022]

Tucker, R., 2009. The anticipatory regulation of performance: the physiological basis for pacing strategies and the development of a perception-based model for exercise performance. *British journal of sports medicine,* 43(6), pp.392-400.

Urquhart, L.D., Craigon, P.J., 2021. The Moral-IT Deck: a tool for ethics by design. *Journal of Responsible Innovation*, 8(1), pp.94–126. doi.org/10.1080/23299460.2021.1880112

Vagia, M., Transeth, A.A., Fjerdingen, S.A., 2016. A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? *Applied Ergonomics*, 53, pp.190–202. doi.org/10.1016/j.apergo.2015.09.013

Vaughan, D., 1997. *The Challenger Launch Decision, The Challenger Launch Decision*. University of Chicago Press, Chicago.

Venables, P.H. and Christie, M., 1980. 'Electrodermal activity', *Techniques in Psychophysiology*, 54(3).

Venkatesh, V. and Davis, F.D., 2000. A Theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management Science*, 46(2), pp.186-204. doi.org/10.1287/mnsc.46.2.186.11926

Warm, J.S., Dember, W.N., Hancock, P.A., 1996. 'Vigilance and Workload in Automated Systems', in Parasuraman, R., Mouloua, M. (eds.), *Automation and Human Performance: Theory and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates. pp.183–200.

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

West, S.W., Williams, S., Kemp, S., Cross, M.J. and Stokes, K.A., 2019. Athlete Monitoring in Rugby Union: Is Heterogeneity in Data Capture Holding Us Back? *Sports*, 7(5), p.98.

Whitlock, A., Pethick, J., Mills, A., 2018. 'Driver Vigilance Devices', in Whitlock, A., Pethick, J. Mills, A. (eds) *Rail Human Factors.* Abingdon, UK: Routledge. doi.org/10.4324/9781315089201-12

Wickens, C. D.2008. "Multiple Resources and Mental Workload." *Human Factors,* 50 (3), pp.449–455.

Wikipedia, 2020. GSM-R. https://en.wikipedia.org/wiki/GSM-R#:~:text=The%20first%20train%20(390%2C034%20on,outside%20of%20the%20Strathclyde%20trial. [Accessed 19.08.2021]

Wilson, J.R., 2014. Fundamentals of systems ergonomics/human factors. *Applied Ergonomics,* 45(1), pp.5–13.

doi.org/10.1016/j.apergo.2013.03.021
Wilson, J.R. and Sharples, S., 2015. 'Methods in the understanding of human factors' in Wilson, J., Sharples, S. (eds.) *Evaluation of Human Work*. 4th Edition. London: CRC Press, pp.1-32.
Wolf, P., Menzel, F., Rennhak, C., 2018. *An Extension of the Technology Acceptance Model tailored to Wearavble Device Technology*. Munich Business School Working Paper Series.
Xie, B. and Salvendy, G., 2000. Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments. *Work Stress,* 14, pp.74–99. doi.org/10.1080/026783700417249
Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A., 2015. State of science: mental workload in ergonomics. *Ergonomics*, 58(1), pp.1-17. doi.org/10.1080/00140139.2014.956151
Young, M.S. and Stanton, N.A., 2002. Malleable attentional resources theory: A new explanation for the effects of mental underload on performance. *Human Factors*, 44(3), pp.365-375. doi.org/10.1518/0018720024497709
Young, M.S. and Steel, T., 2017. Rail worker fatigue: Identification, management and countermeasures. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, 231(10), pp.1098–1106. doi.org/10.1177/0954409716675383
Zawiah, N.I.A.Rahman, and Dawal, S.Z.M., 2016. The Mental Workload and Alertness Levels of Train Drivers Under Simulated Conditions Based on Electroencephalogram. *Malaysian Journal of Public Health Medicine.* pp.115-123.
Zeilstra, M., 2021. The Network Rail Signaller Task Demand Tool: theoretical background and preliminary tool. *Seventh International Human Factors Rail Conference*. 23-25 June 2021. Rail Safety and Standards Board. Live virtual conference, https://www.ukrrin.org.uk/events-and-activity/the-seventh-international-rail-human-factors-rail-conference/ [Accessed 09.02.2023]

Appendix A. Interviews and familiarisation ethics and prompts

Ethics for Industry Interviews

Ethics Application - Abigail Fowler - Decision

EZ-Eng-Ethics <EZ-Eng-Ethics@exmail.nottingham.ac.uk>
Thu 20/07/2017 16:53
To: Abigail Fowler <psxacf@exmail.nottingham.ac.uk>

1 attachments (28 KB)
Reviewer Decision_DEM.docx;

Hi Abigail,

Please find attached the final decision on your recent ethics application.

The decision is: Approval Awarded – no changes required.

Best of luck with your study.

Many Thanks,

Research Policy and Governance Officer

Faculty of Engineering
University of Nottingham
B03, L4
University Park
Nottingham, NG7 2RD

+44 (0) 115 95 15561 I donna.astill-shipman@nottingham.ac.uk

Part 1 Challenges in Rail related to human staff performance DATE: REF:

- **WHAT** – Current **challenges** in rail that rely on **human performance**
- **WHAT – New technology** involved
- **WHO – Staff** roles involved
- **WHY** are the above challenges? *(e.g. ' performance, safety, cost' ?)*
- **HOW** - does challenge vary e.g. Day/night, weather, type of ops.
- **WHERE –** does it vary geographically, or across different rail operations?
- **Impact –** likelihood, risk level and how is impacted, location?
- **Rank** the challenges in order of priority for rail, highest first. *Why?*

Part 2 Data on Human Performance

- **WHAT** – **data** currently exists in rail that captures human performance (*e.g. experience, incidents*)
- **WHY** is this data collected?
- **WHAT technology** is currently used to collect data?
- **HOW** is the data **collected?**
- **WHO uses** data collected on human performance
- **Wearable** use in rail or activity apps?

Part 3 Other Perspectives in Rail - Who to interview?

WHO – name/role/organization + can I say you recommended them?

Ethics for Industry Familiarisation

Ethics Application - Decision - Human performance in rail: understanding signalling operations

EZ-Eng-Ethics
Tue 08/01/2019 11:45
To: Abigail Fowler <psxacf@exmail.nottingham.ac.uk>
Cc: Catherine Harvey <ezzch1@exmail.nottingham.ac.uk>

1 attachments (28 KB)
Reviewer Decision_M03.docx;

Dear Abi,

Please find attached the final decision on your recent ethics application.

The decision is: Approval Awarded – no changes required.

Best of luck with your study.

Best wishes,

Emily Judd
Research Administrator

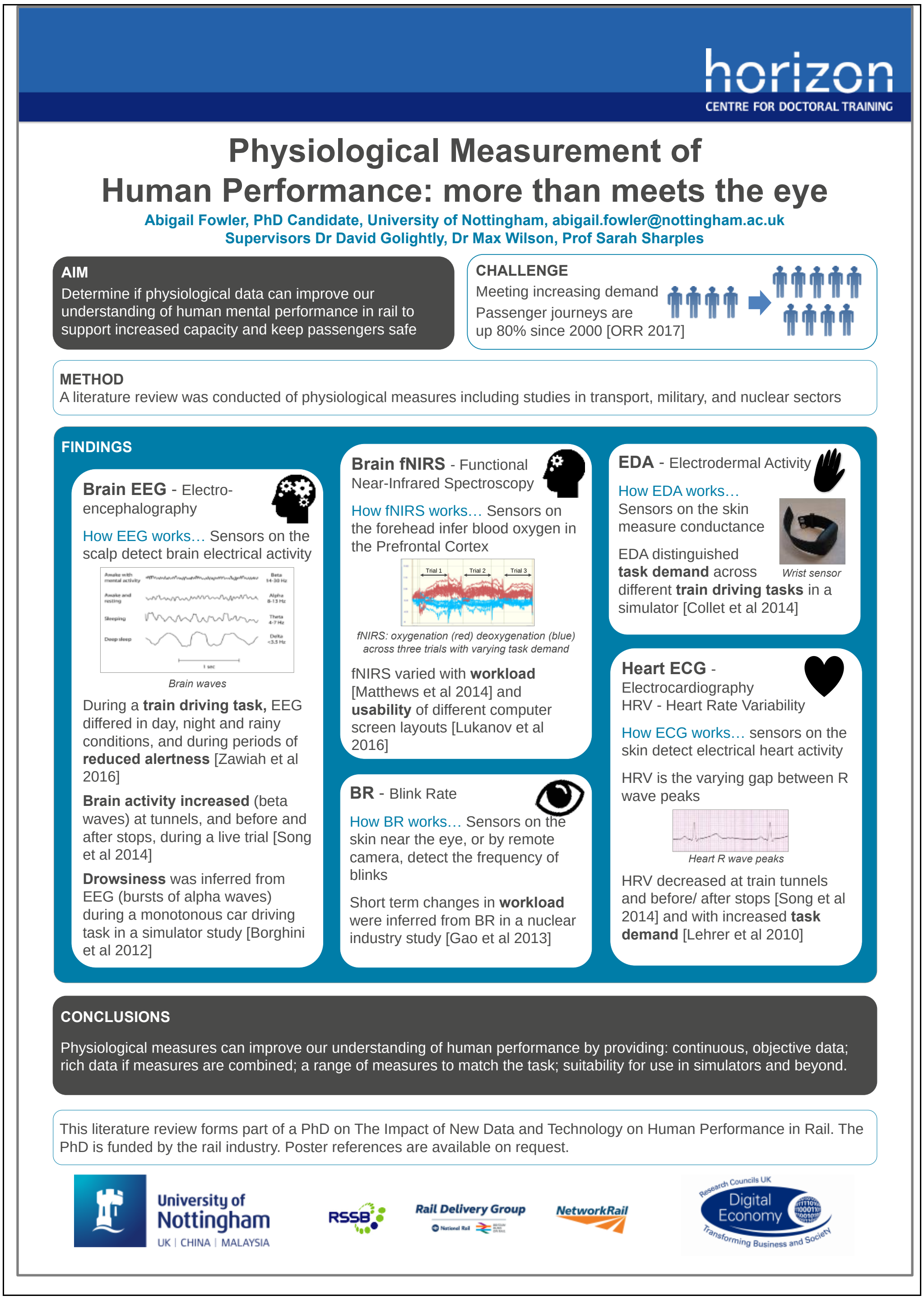
Faculty of Engineering
University of Nottingham
APM Hub, B03, L4, Coates Building
University Park
Nottingham, NG7 2RD

Research Groups: OAP, Environmental Fluids, GGIEMR and Resilience Engineering
Research Priority Area: Healthcare Technologies

Emily.Judd@nottingham.ac.uk
0115 74 84495
Hours of work: Monday-Wednesday & Friday 8.30am to 12:00am. Thursday 8:00am to 12:30pm.
Please consider the environment before printing off this email

Appendix B. Physiological measurement poster

This poster was displayed at the HFRail Conference, London, November 2017.



References from poster: (Borghini et al., 2012; Collet et al., 2014; Gao et al., 2013; Lehrer et al., 2010; Matthews et al., 2014; Y. Song et al., 2014; Zawiah and Dawal, 2016)

Appendix C. Simulation study ethics, NTS and storyboards

Ethics for Simulation Study

Ethics Application - Decision - Using wearable physiological devices to measure

EZ-Eng-Ethics <EZ-Eng-Ethics@exmail.nottingham.ac.uk>
Wed 17/04/2019 08:32
To: Abigail Fowler [redacted]

2 attachments (61 KB)
Reviewer Decision_DE13.docx; Reviewer Decision_R08.docx;

Dear Abigail,

Please find attached the final decision on your recent ethics application.

The decision is: Approval Awarded – no changes required.

Best of luck with your study.

Best wishes,

[redacted]
Research Administrator

Faculty of Engineering
University of Nottingham
APM Hub, B03, L4, Coates Building
University Park
Nottingham, NG7 2RD

Research Groups: OAP, Environmental Fluids, GGIEMR and Resilience Engineering
Research Priority Area: Healthcare Technologies

Emily.Judd@nottingham.ac.uk
0115 74 84495
Hours of work: Monday-Wednesday & Friday 8.30am to 12:00am. Thursday 8:00am to 12:30pm.
Please consider the environment before printing off this email

Non-Technical Skills

These are the Non-Technical Skills used in the rail industry in the UK (RSSB 2012, RSSB 2016).

Table 1 - GB Rail Non-Technical Skills

NTS categories	Skill
1. Situational awareness	1.1 Attention to detail 1.2 Overall awareness 1.3 Maintain concentration 1.4 Retain information 1.5 Anticipation of risk
2. Conscientiousness	2.1 Systematic and thorough approach 2.2 Checking 2.3 Positive attitude towards rules and procedures
3. Communication	3.1 Listening 3.2 Clarity 3.3 Assertiveness 3.4 Sharing information
4. Decision making and action	4.1 Effective decisions 4.2 Timely decisions 4.3 Diagnosing and solving problems
5. Co-operation and working with others	5.1 Considering others' needs 5.2 Supporting others 5.3 Treating others with respect 5.4 Dealing with conflict or aggressive behaviour
6. Workload management	6.1 Multi-tasking and selective attention 6.2 Prioritising 6.3 Calm under pressure
7. Self-management	7.1 Motivation 7.2 Confidence and initiative 7.3 Maintain and develop skills and knowledge 7.4 Prepared and organised

References

RSSB (2012) Non-Technical Skills http://www.rssb.co.uk/Library/improving-industry-performance/2012-leaflet-non-technical-skills.pdf [Accessed 18/09/2022]

RSSB (2016) A Good Practice Guide to Integrating Non-Technical Skills into Rail Safety Critical Roles. https://www.rssb.co.uk/Library/improving-industry-performance/2016-07-non-technical-skills-integration-good-practice-guide.pdf [Accessed 17.09.18]

Storyboards

Storyboards were created in the simulation study from each Individual participants' temporal task demand, events, EDA, HRV and IWS data. All nineteen participants are presented here (P7 was excluded due to a technical problem). Where HRV data, or IWS data points are missing there are not estimated, they are omitted from the graphs.

The top graph plots: Task Demand (Total trains including Freight), with the freight train indicated in a darker grey; EDA (µx100); and wrist accelerometer data.

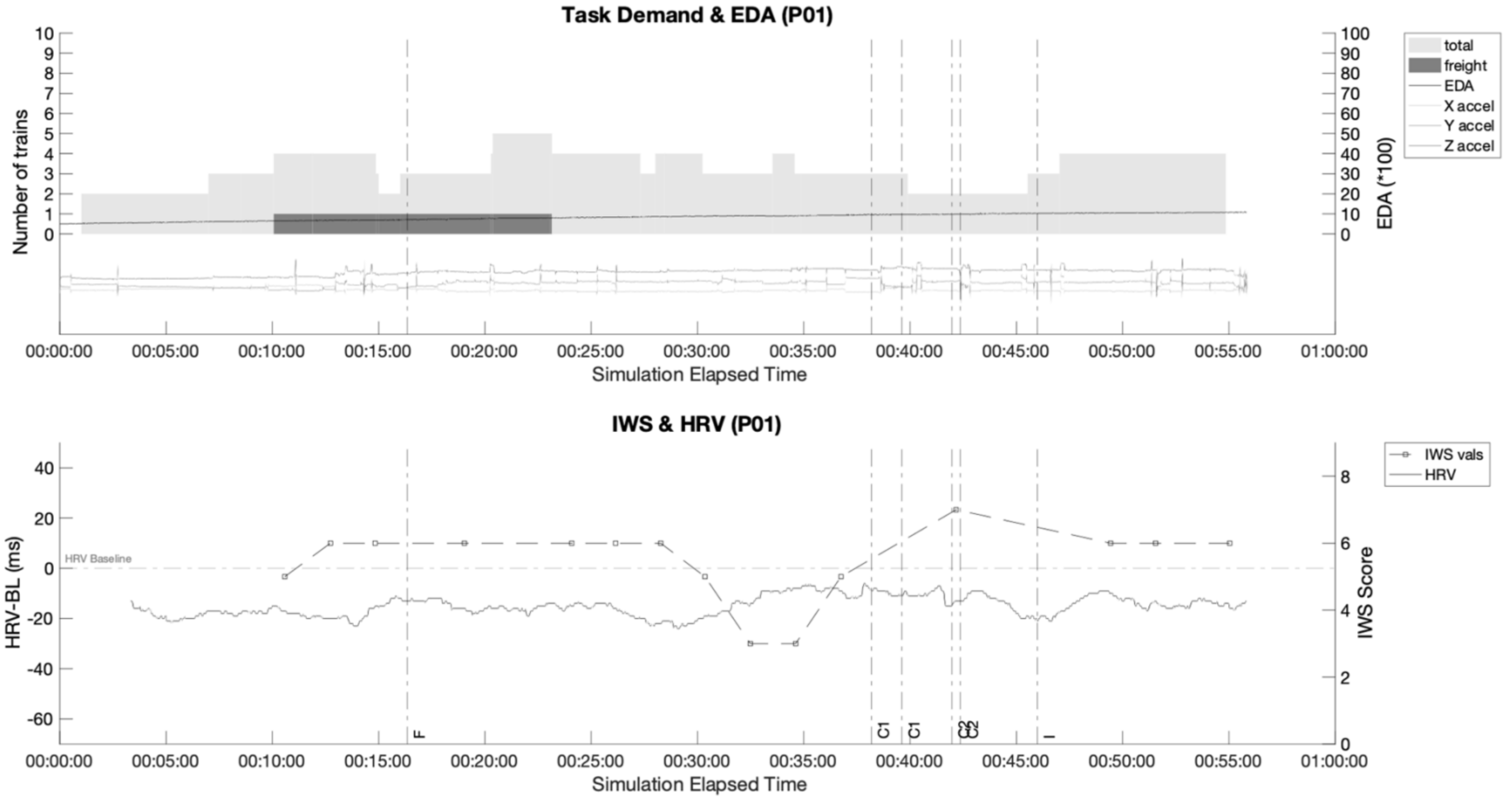
The bottom graph plots: IWS rating; and HRV (ms).

Task Events are marked vertically:

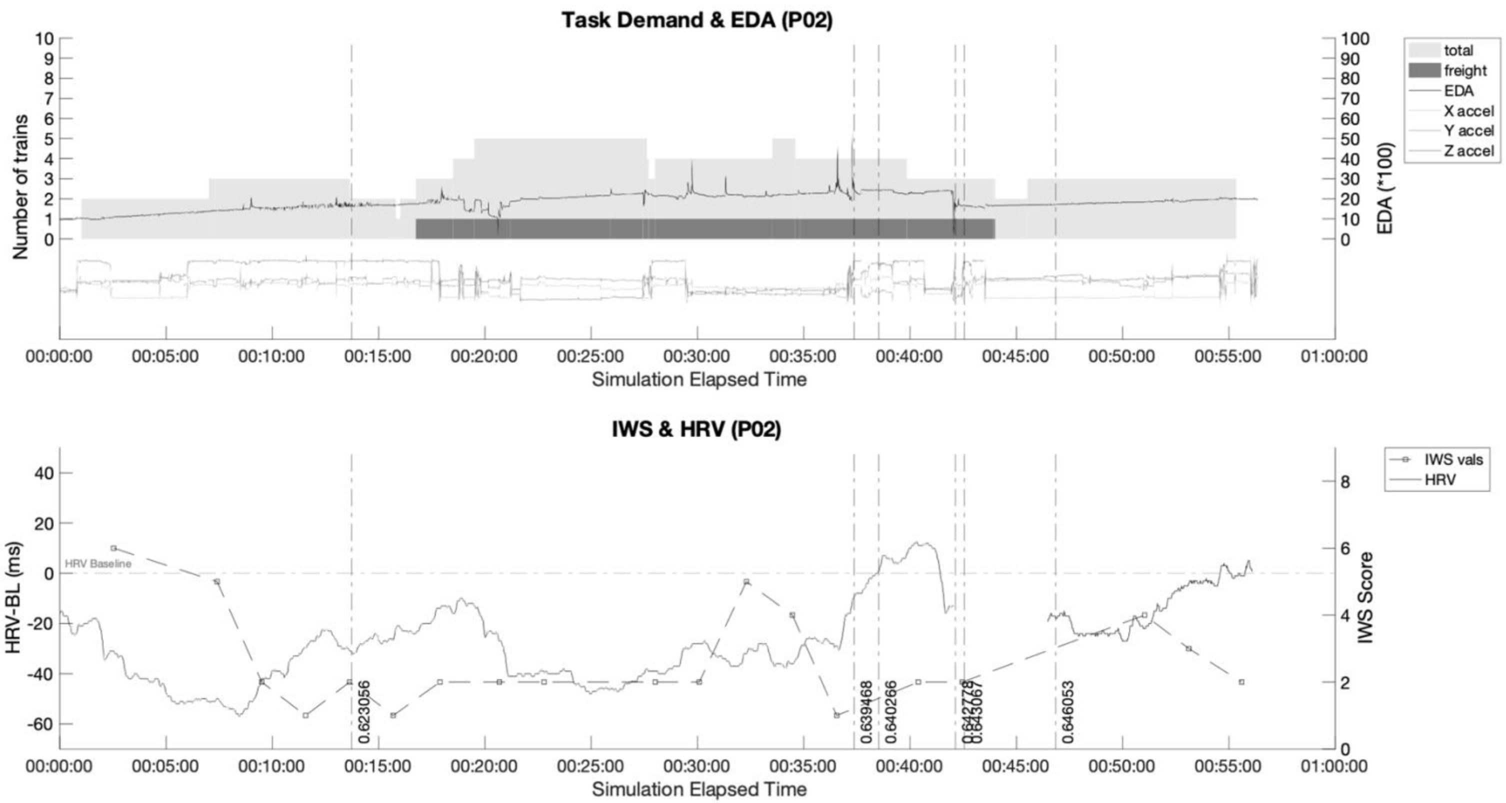
I = Interpose headcode (the train reporting number)

C = call start and end

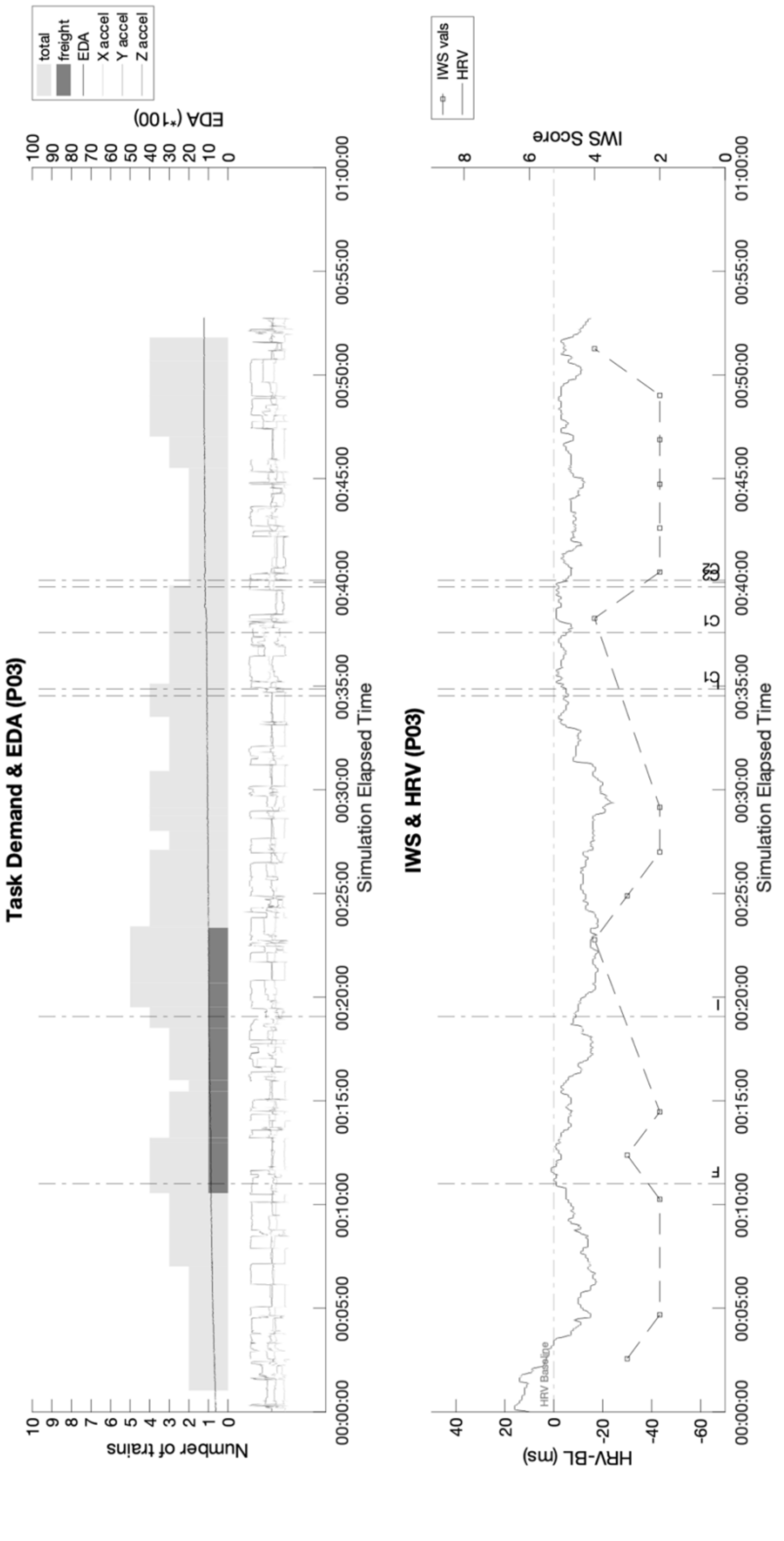
F = Notices freight.



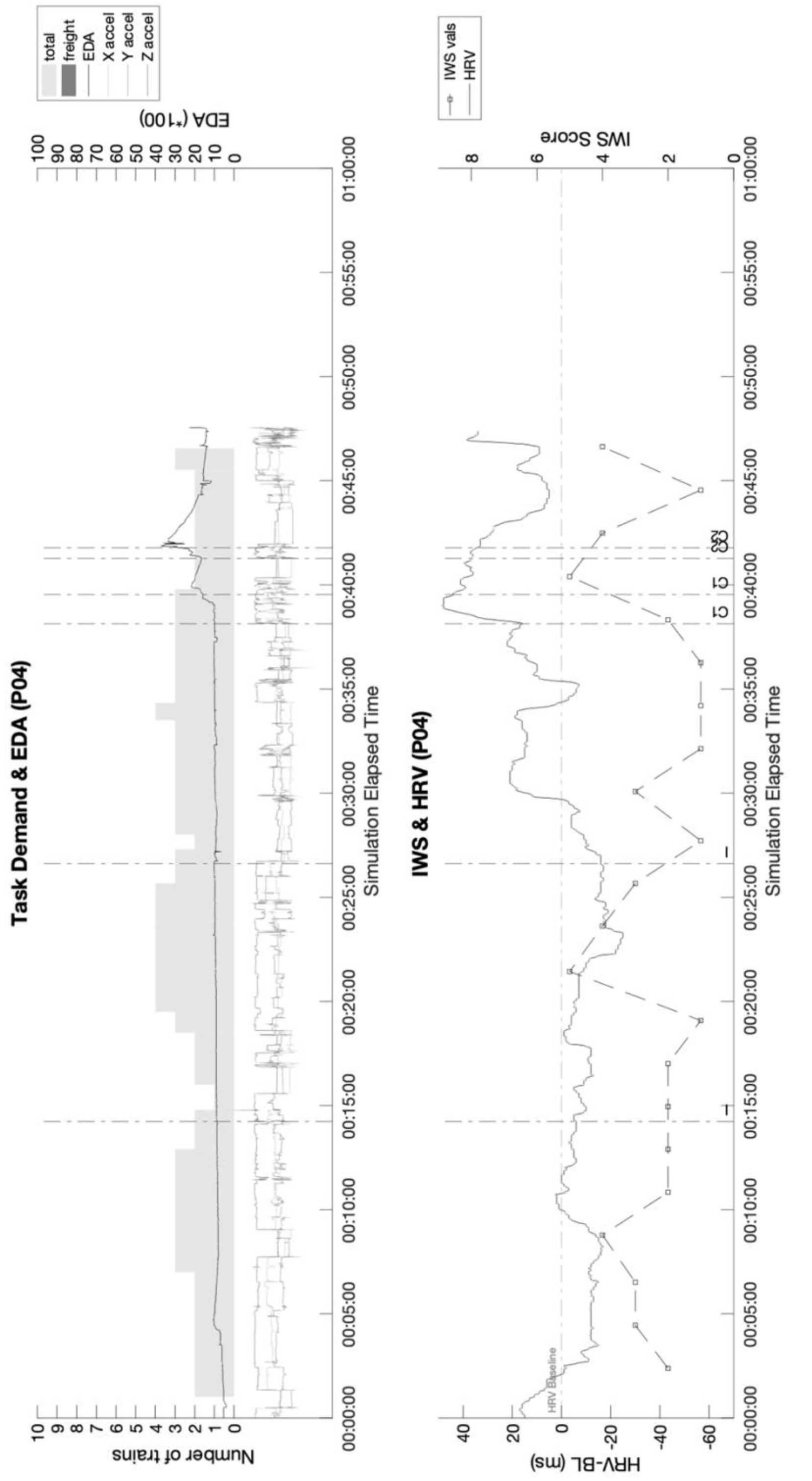
Storyboard (P1). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



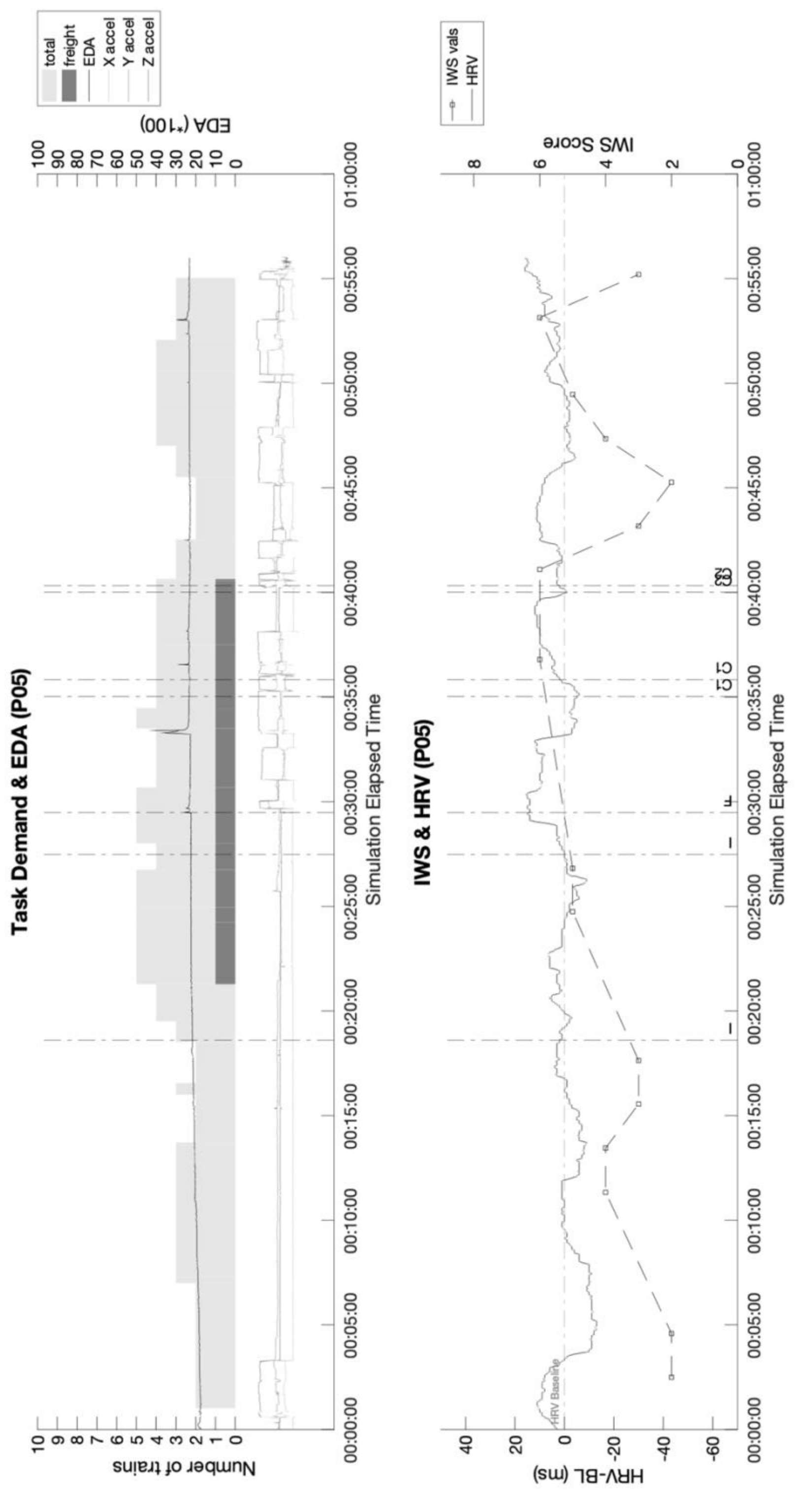
Storyboard (P2). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



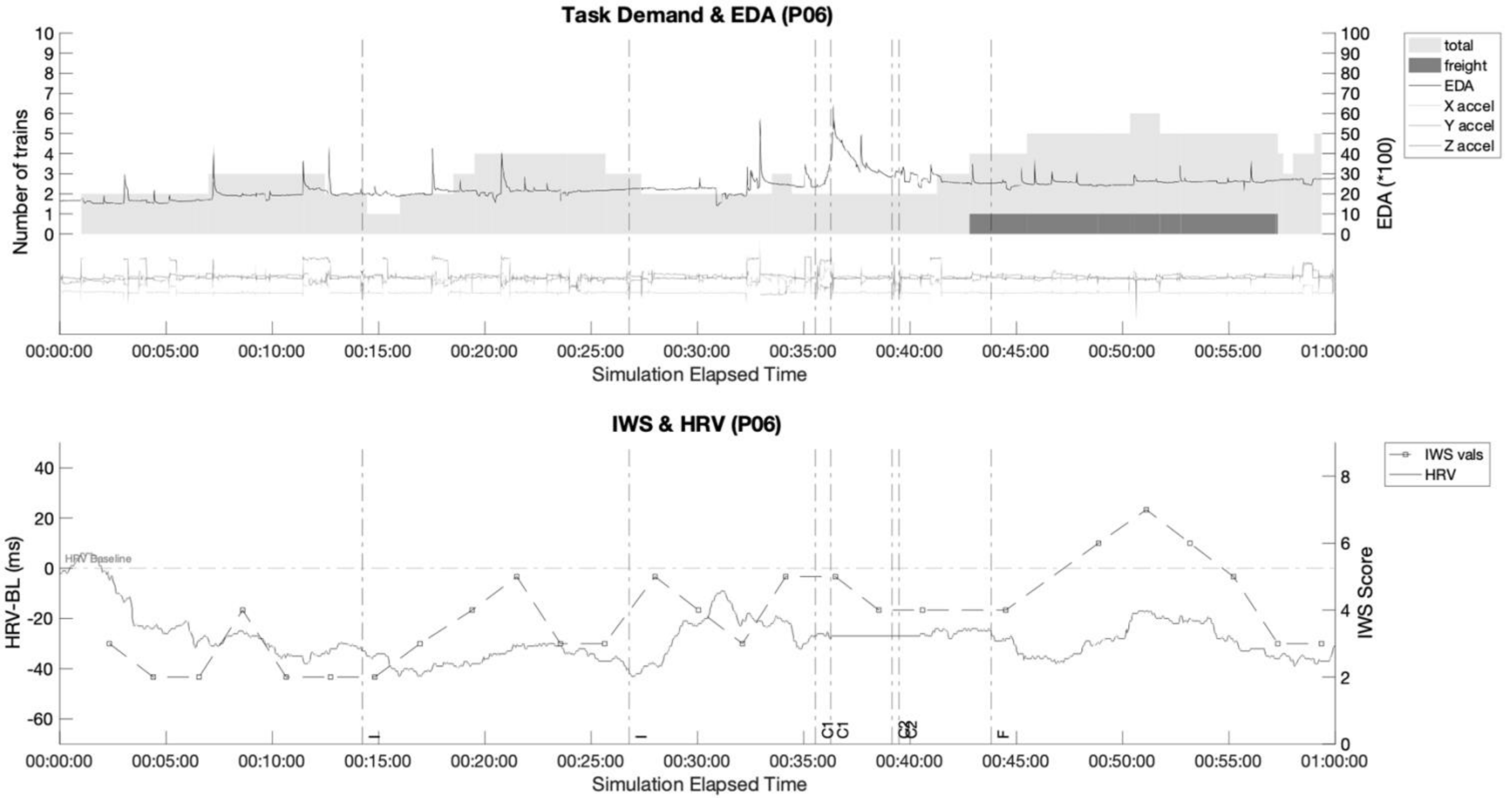
Storyboard (P3). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



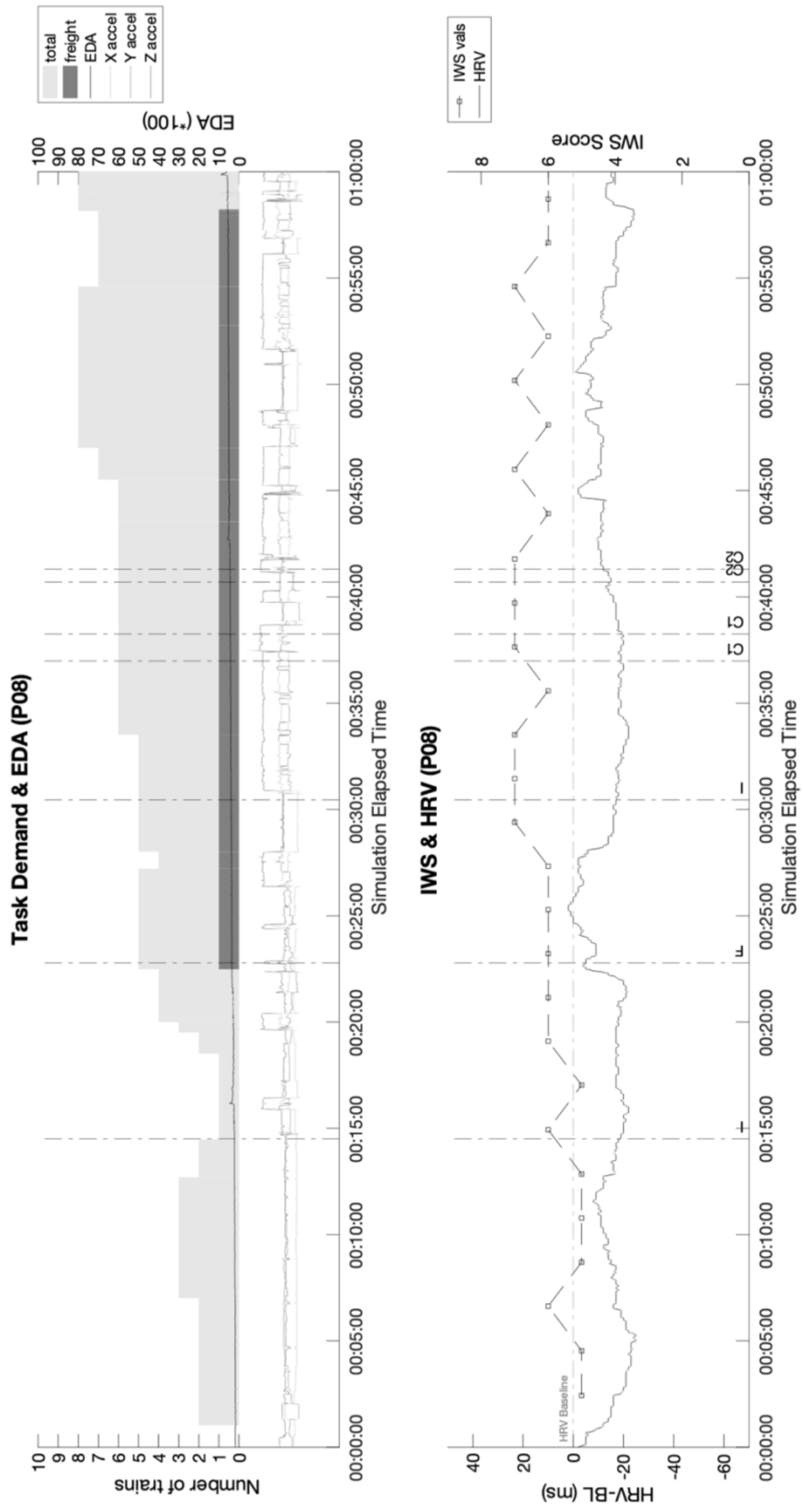
Storyboard (P4). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



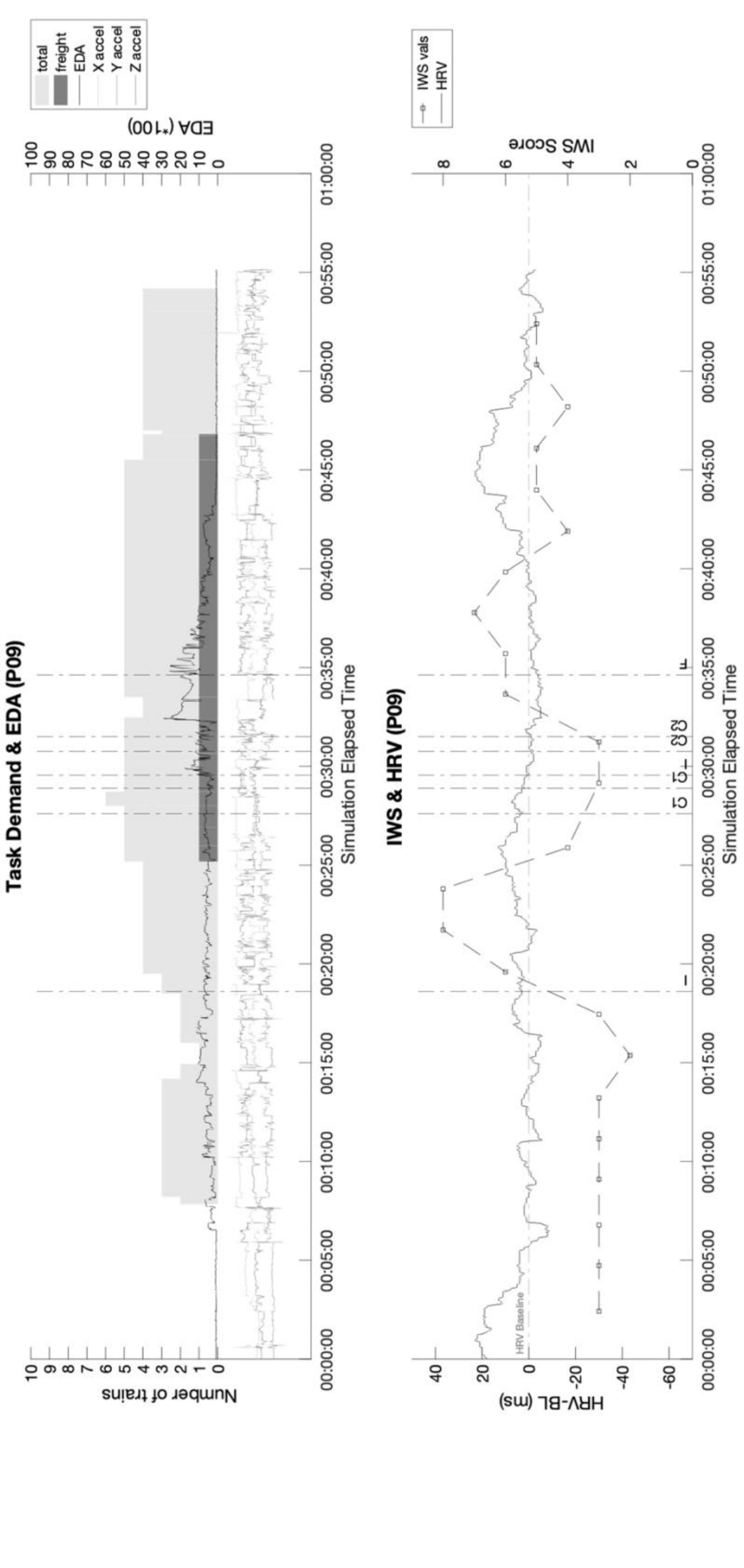
Storyboard (P5). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



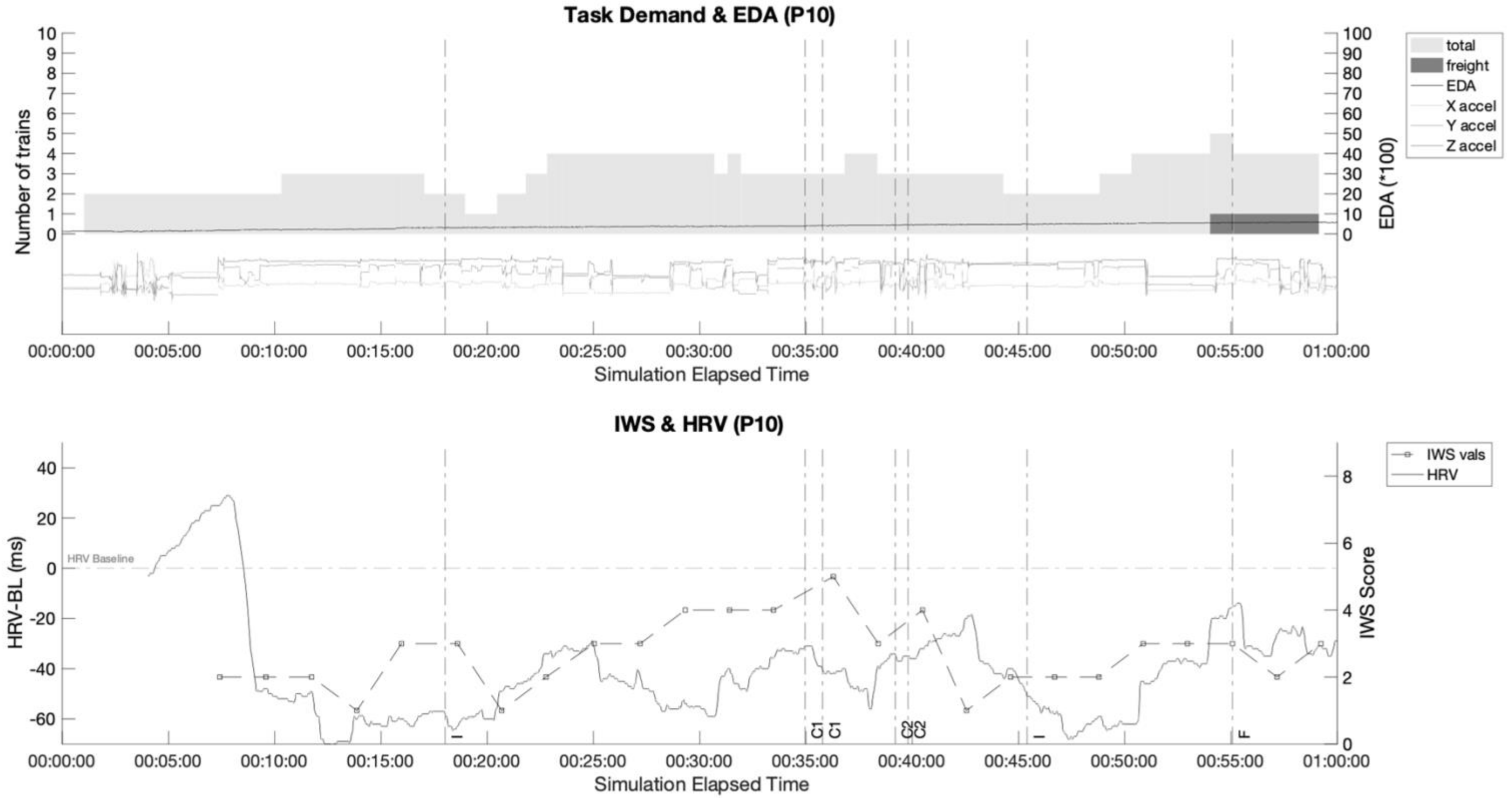
Storyboard (P6). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



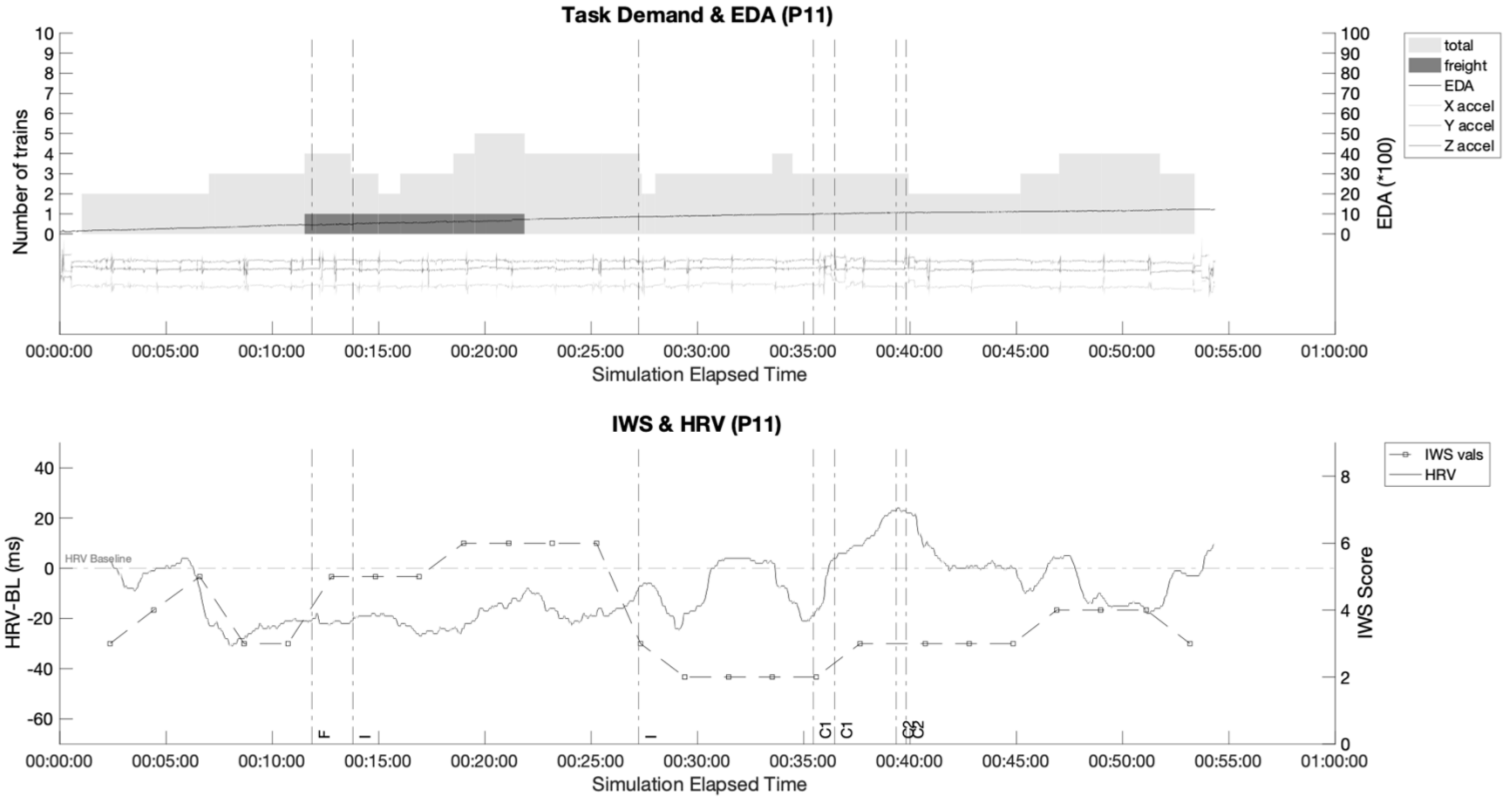
Storyboard (P8). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



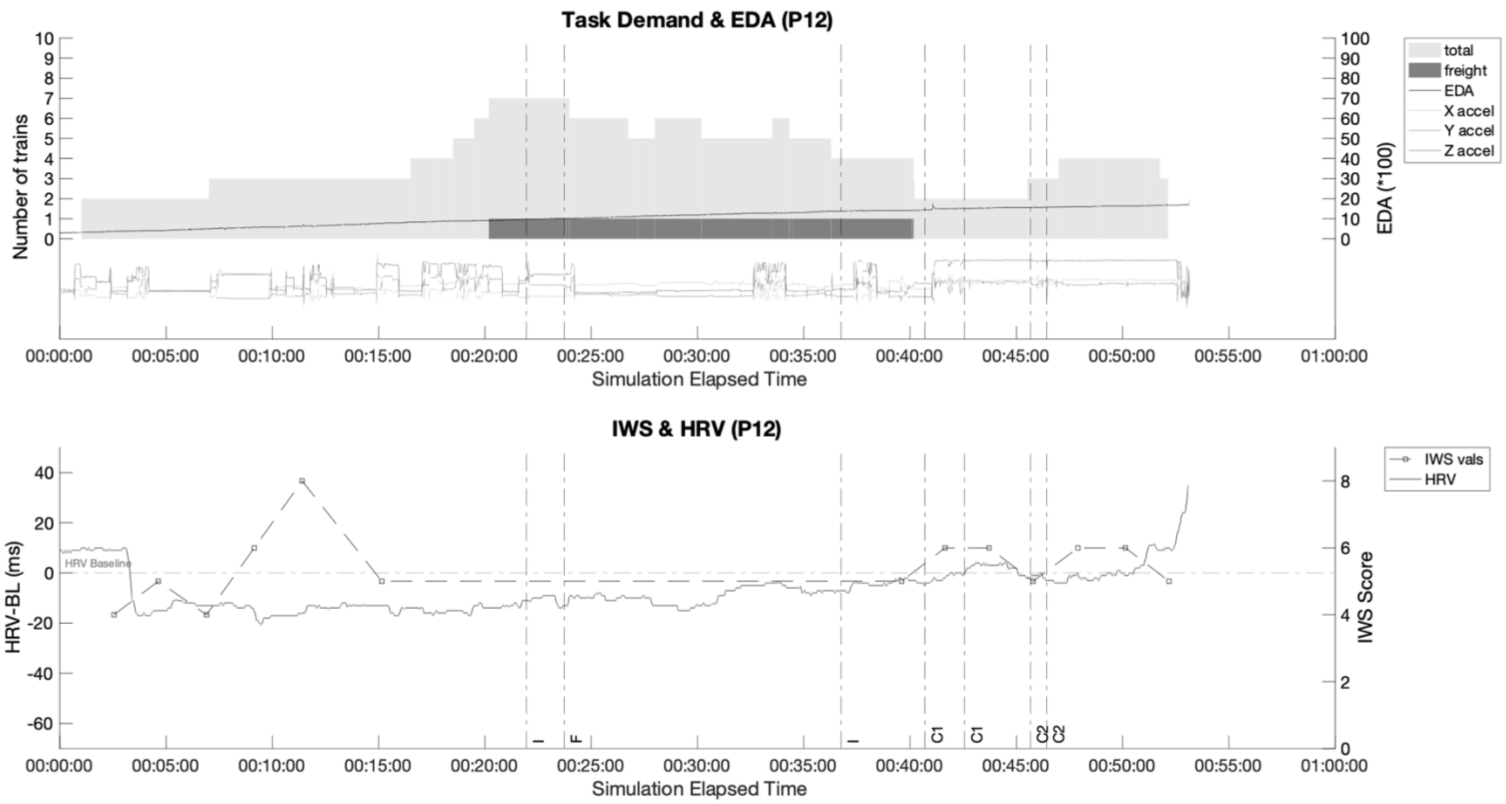
Storyboard (P9). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



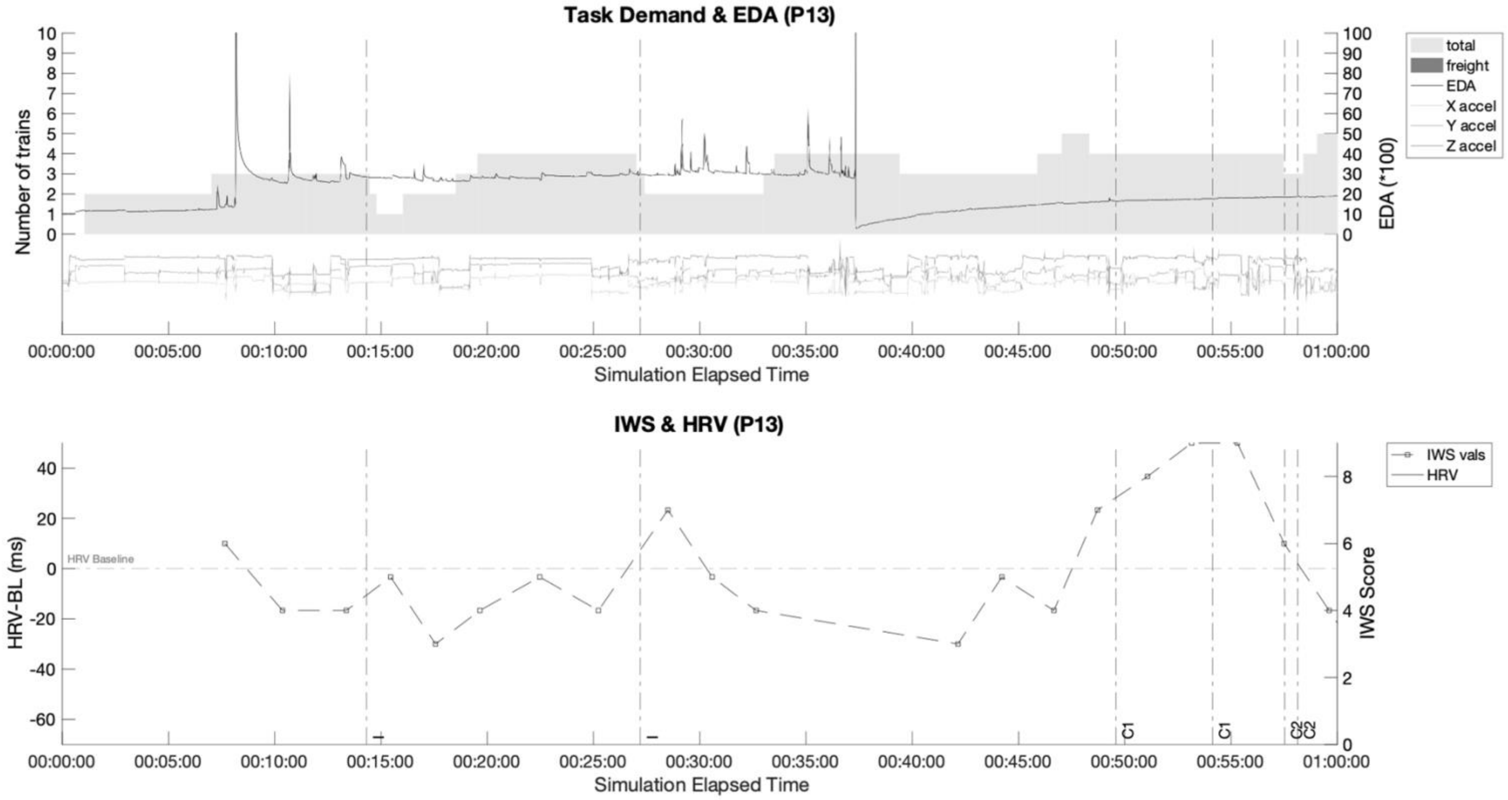
Storyboard (P10). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



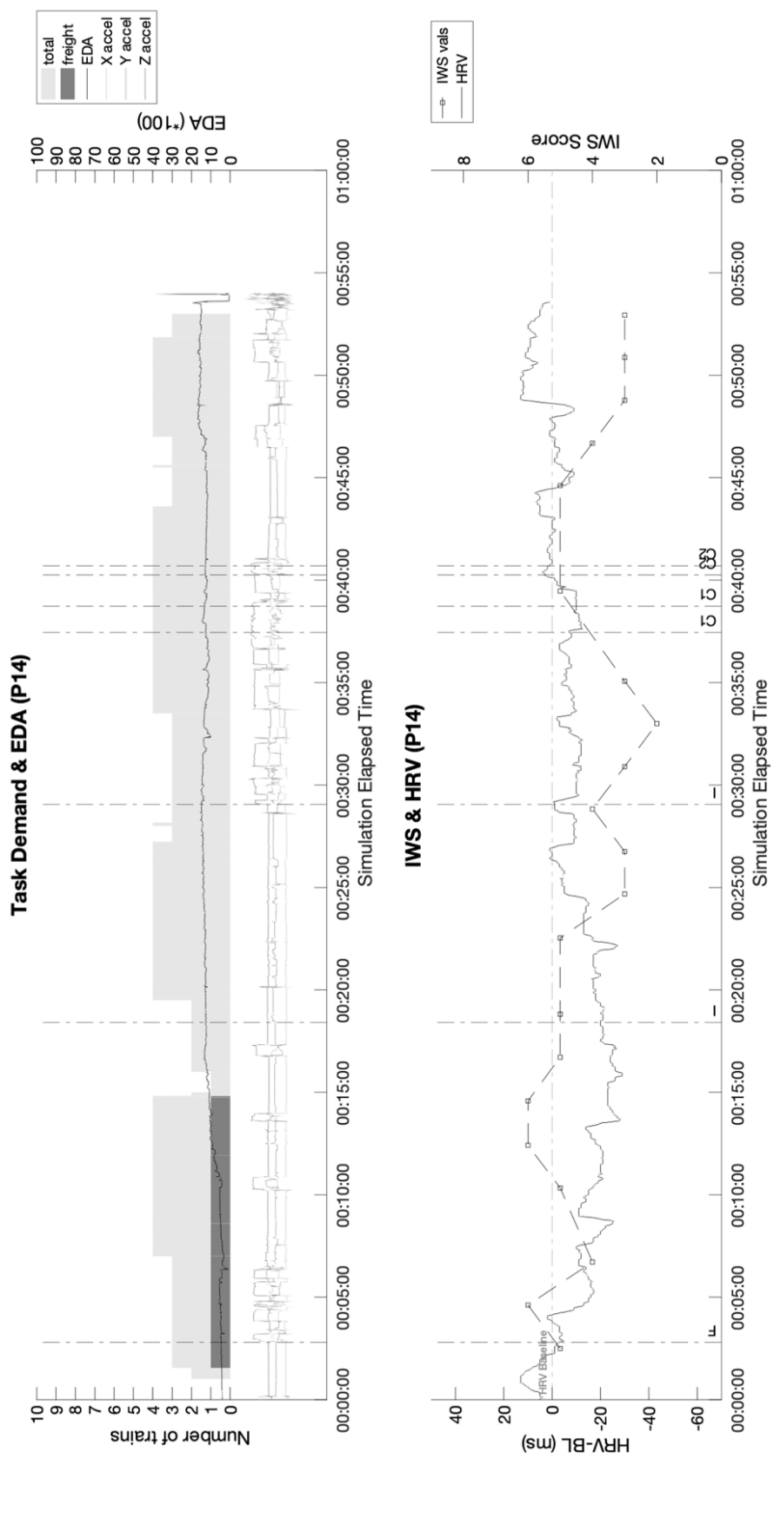
Storyboard (P11). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



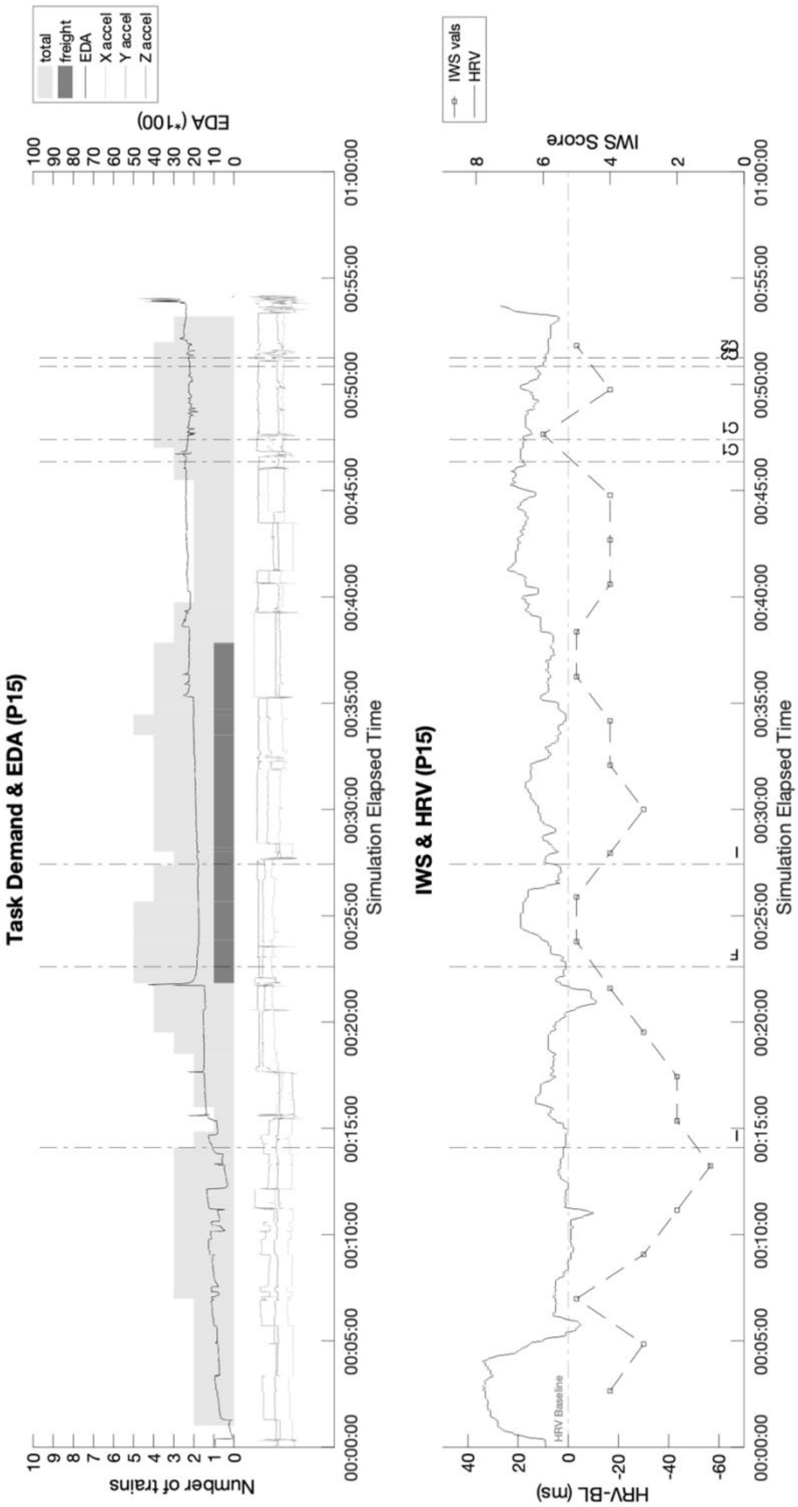
Storyboard (P12). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



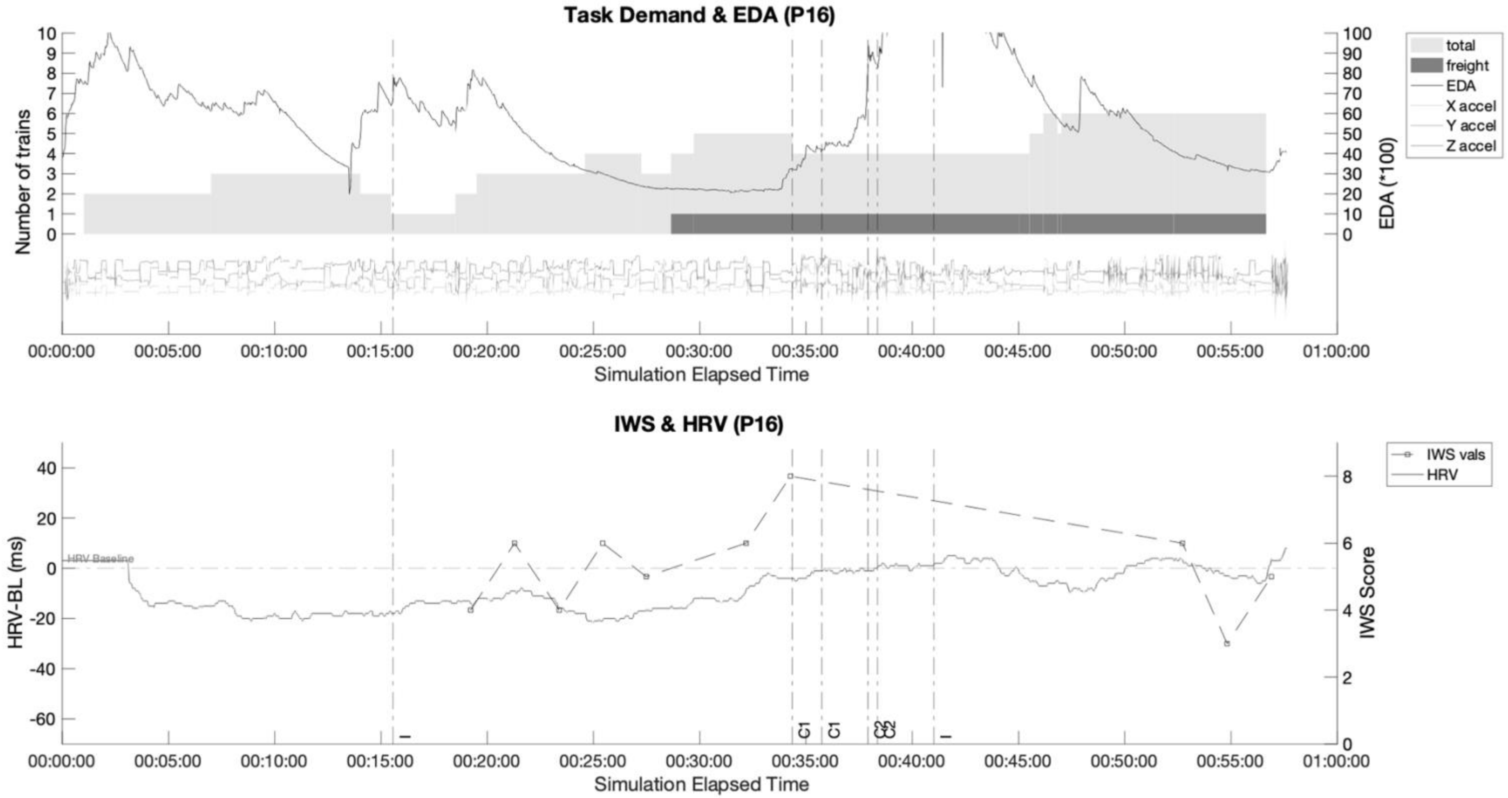
Storyboard (P13). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



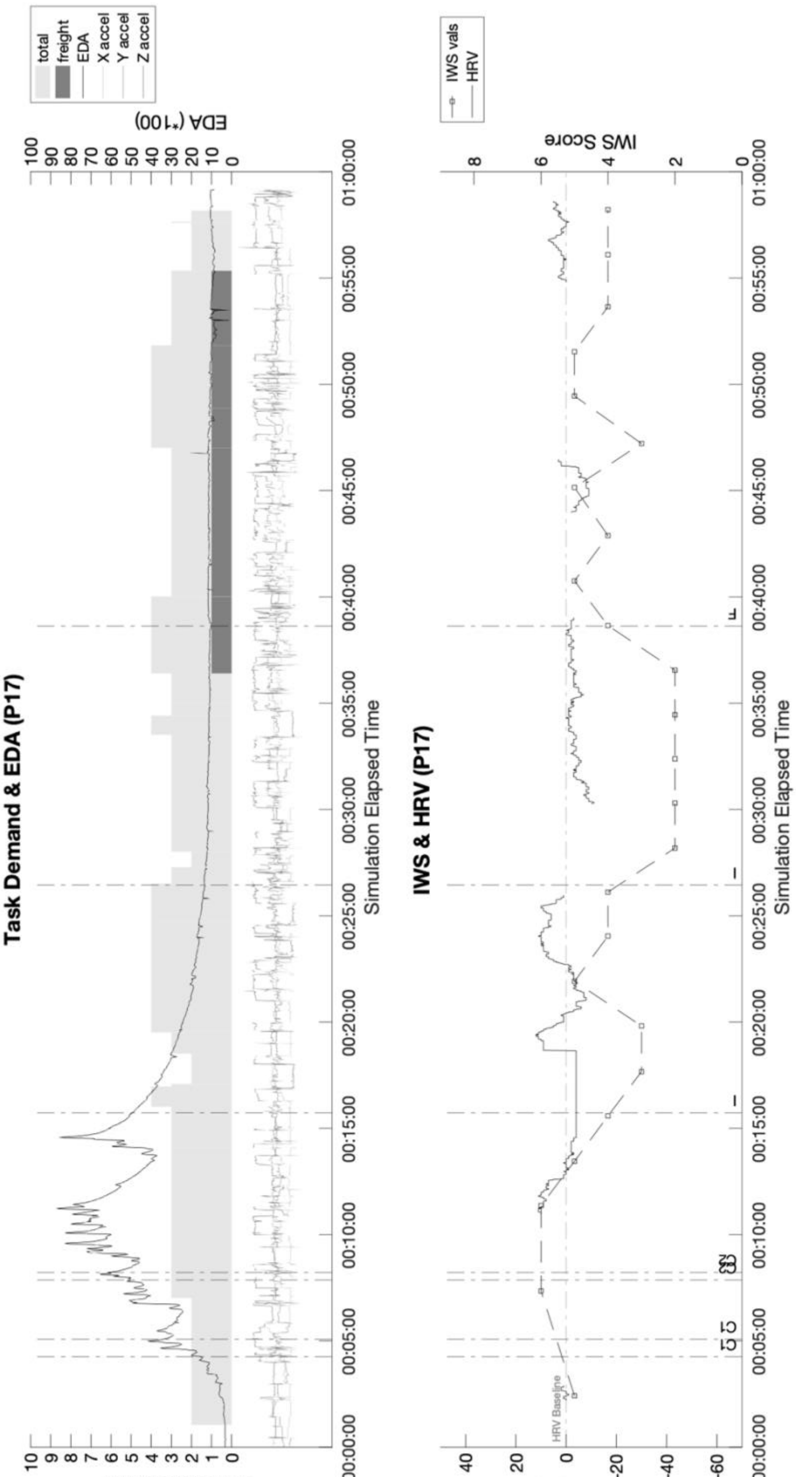
Storyboard (P14). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



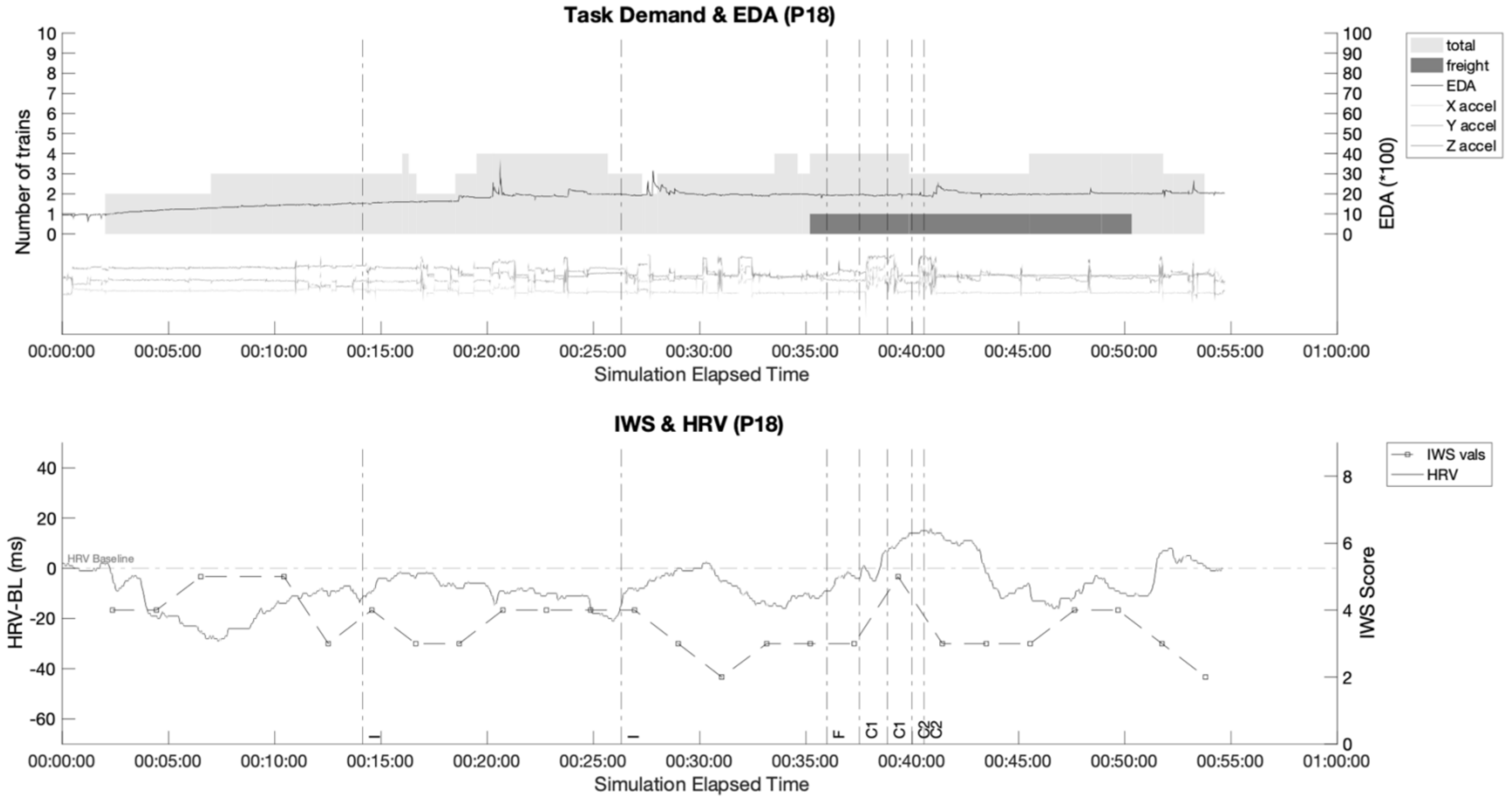
Storyboard (P15). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



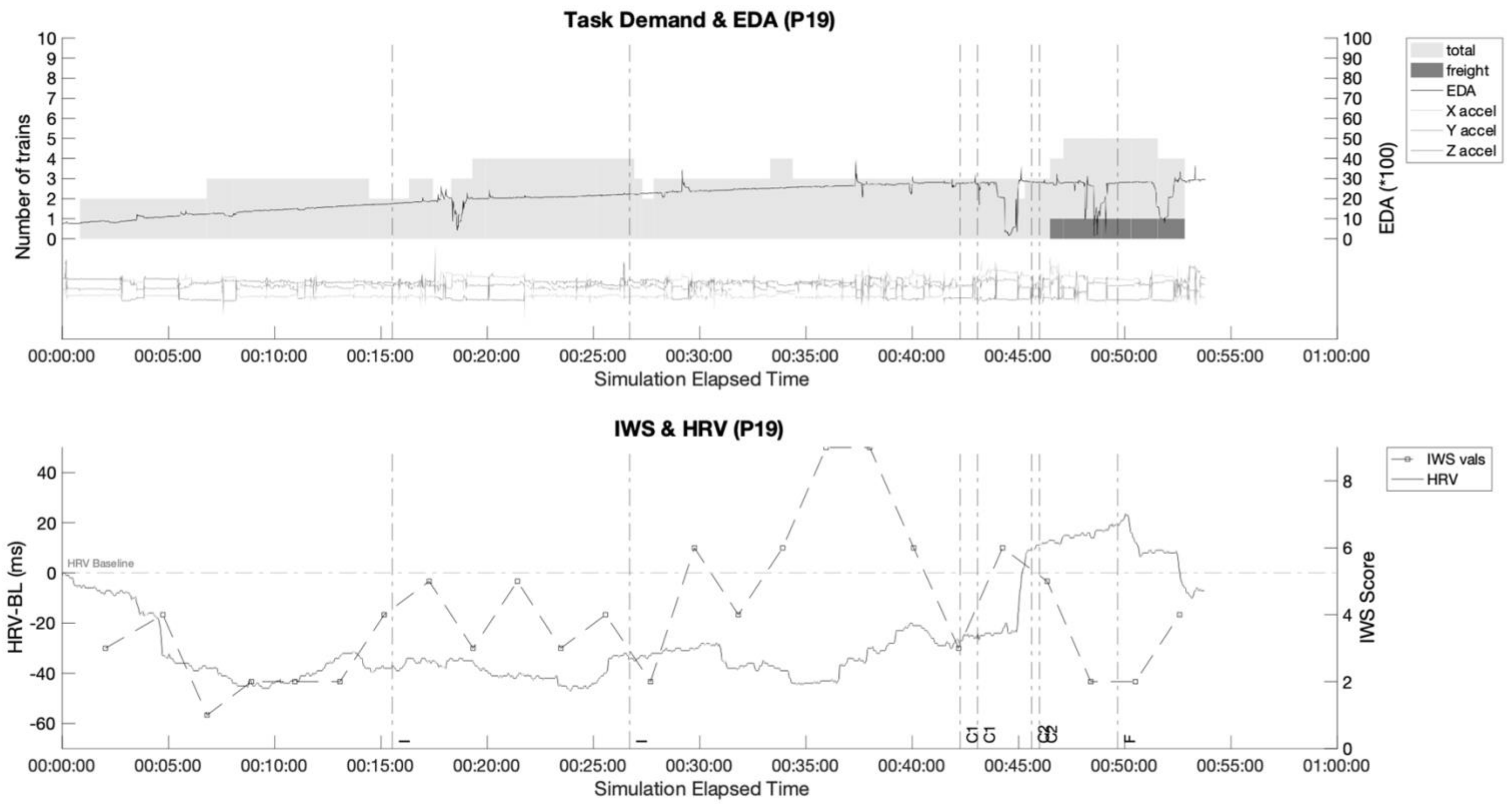
Storyboard (P16). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



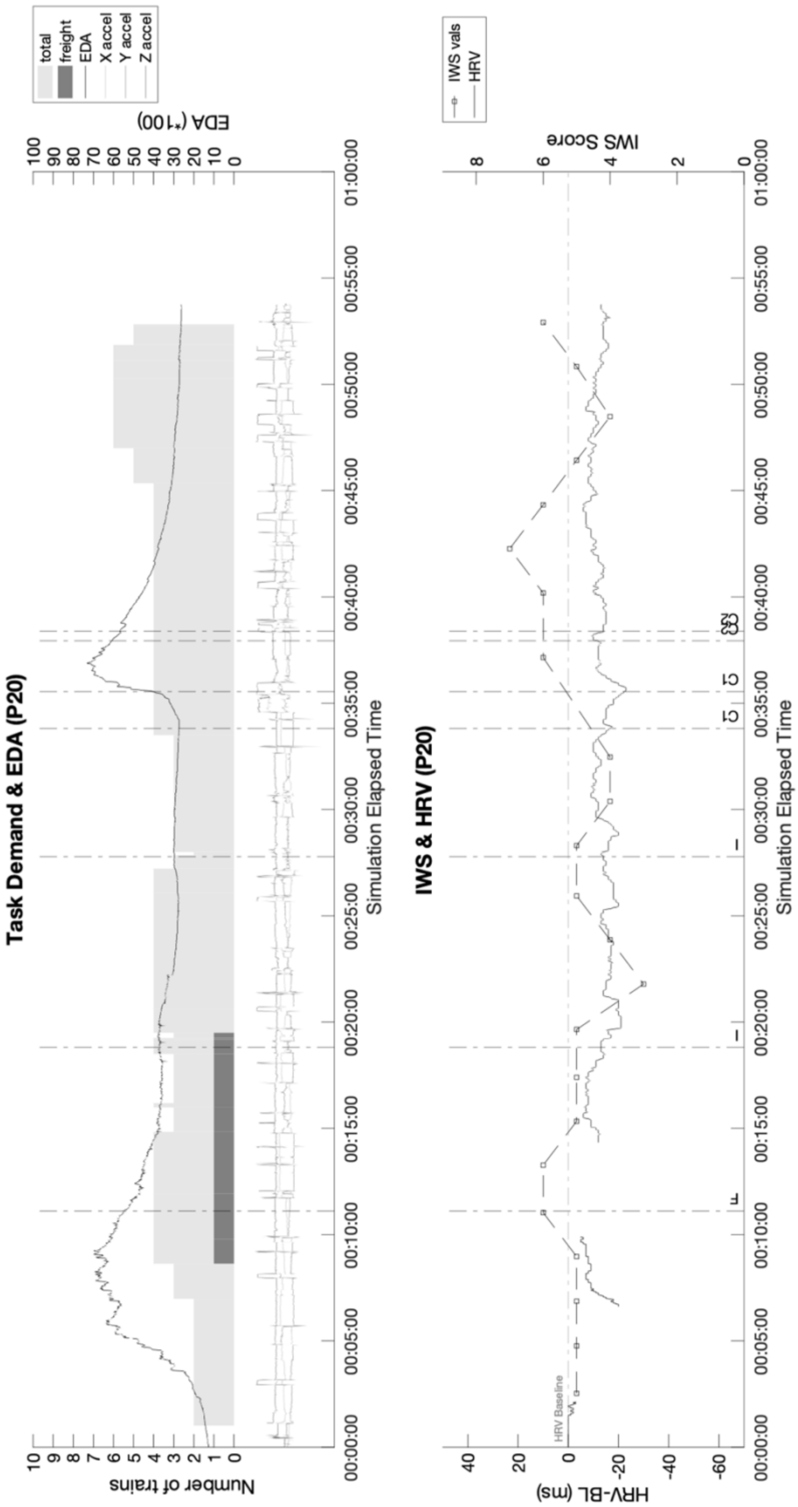
Storyboard (P17). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



Storyboard (P18). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



Storyboard (P19). Top graph: Task demand (Total trains including Freight); EDA (μSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).



Storyboard (P20). Top graph: Task demand (Total trains including Freight); EDA (µSx100); and wrist accelerometer data. Bottom graph: IWS rating; and HRV (ms). Task Events marked vertically (I = Interpose headcode, C = call start and end, F = Notices freight).

Appendix D. Attitudes study ethics and theories

Ethics for Attitudes Study

Ethics Application - Decision - Attitudes to Wearables in the Workplace

EZ-Eng-Ethics <EZ-Eng-Ethics@exmail.nottingham.ac.uk>
Tue 29/09/2020 10:05
To: Abigail Fowler [redacted]

1 attachments (39 KB)
Reviewer Decision_818.docx;

Hi Abi,

Please find attached the final decision on your recent ethics application.

The decision is: Approval Awarded – no changes required.

Best of luck with your study.

Many Thanks,

[redacted]

Senior Research Operations Officer

In line with UK Government advice on Covid-19, the University of Nottingham has closed many of its non-essential buildings and I am currently working from home. Apologies there may be a delay in my response to your emails.

Faculty of Engineering
University of Nottingham
B03, L4
University Park
Nottingham, NG7 2RD

+44 (0) 115 95 15561 I donna.astill-shipman@nottingham.ac.uk

Hours of work: Monday, Tuesday, Wednesday, Friday – 8:15am – 4:30pm
Thursday – 8:15am – 2:30pm

Engineering Research Handbook

Mapping theory to topics

KEY:

G Gribel, Regier and Stengel (2016)

J Jacobs *et al.* (2019)

P Parasuraman and Colby (2015)

U Urquhart and Craigon (2020)

V Venkatesh and Davis (2000)

W Wolf *et al.* (2018)

Pre-Interview Questionnaire			
	Topic	**Question**	**Adapted from**
1	Experience of signalling	Date first passed competent as a signaller Date first passed competent on a VDU workstation	New item
3	Experience of workload assessment	Have you ever taken part in a workload assessment at work? • You carried out work whilst someone observed for a workload assessment • You carried out work and rated your workload on a scale 1-9 every few minutes • You rated workload on 3 scales after you finished your tasks • You were asked to count the number of items or features on a workstation • You listed tasks in a typical hour at a new workstation, in a workshop	New item
2	Personal experience of wearables	How often do you collect data from wearables or fitness apps? (1 Never, 2 Previously, not now, 3 Sometimes, 4 Weekly, 5 Daily) Wearables: • Step counter • Wrist strap fitness/heart tracker • Chest strap fitness/heart tracker • Smart phone with fitness app • Smart watch with fitness app	G, J, V

Interview Question			
	Topic	**Question**	**Adapted From**
1	Experience	If you use wearables, what's your favourite and why?	J, V, G
2.1	**Devices** Perceived Ease of Use	How distracting could devices be for you during a shift? (chest, wrist, app)	P, G, J, U
2.2		How comfortable could devices be for you during a shift? (chest, wrist, app)	U, W
3	Subjective Norm/ Image	What is the likely reaction of colleagues to the use of these new measures (chest, wrist, app)?	U, V
4	Own experiences of workload	From your experience on shift give examples of: • Anticipating a problem correctly (a 'gut' reaction) • Time pressure/ stress • Unexpected event/ system response not expected • Brief workload peak affected you for rest of shift • Confidence when successful achieving your tasks • Impact of a change of new technology/ procedures	New item
Use Cases			
A.	**Use Case** Understand Signaller Workload	Assess *task* – peaks/troughs in workload & overall effort Devices Infer – anticipation, alertness, stress, time pressure, brief peaks Data – anonymised When – 1-2 shifts	New item New item J, U, W New item
5.1	Perceived Usefulness	How useful would it be to demonstrate to *others how hard you work*?	New item
5.2	Anonymity	Acceptable?	J, U, W
5.3	Job Relevance	What's inferred – which most relevant to assessing *task*?	V
5.4	Trust	Any concerns about data use?	J, P, G
5.5	Time	Data collection needed vs tolerable: >4 hrs? Continuous or every 5 mins (app)	New item
B.	**Learning Aid**	Assess *trainee* – track progress, self-learning & training effectiveness. Infer – alertness, confidence, unexpected events, stress, effort Data – You & trainer share. (Device supplier?) When – Before/during/after training to plot progress	New item New item J, G, U, W New item
6.1	Perceived Usefulness	How useful would it be for *you* to understand your data (e.g. heart)?	New item
6.2	Anonymity	Acceptable?	J, G, U, W
6.3	Job Relevance	What's inferred – which most relevant to assessing *trainees?*	New item
6.4	Trust	Any concerns about data use?	J, P, G
6.5	Time	Data collection needed vs tolerable: >4 hrs? Continuous or every 5 mins (app)	New item
C.	**Assess impact of new tech or procedures**	Assess *change* – work support/ effectiveness of change Infer – stress, effort, unexpected system responses Data – Workstation & initials. Managers may know. Supplier? Investigator? When – Before/during/after change.	New item New item J, G, U, W New item
7.1	Perceived Usefulness	How useful would it be to demonstrate to *others impact of changes* on signallers*?*	New item
7.2	Anonymity	Acceptable?	J, G, U, W
7.3	Job Relevance	What's inferred – which most relevant to assessing impact of *change*?	New item
7.4	Trust	Any concerns about data use?	J, P, G
7.5	Time	Data collection needed vs tolerable: >4 hrs? Continuous or every 5 mins (app)	New item
8	Precautionary Principle	Just because we could use these measures, should we?	U

Post-Interview Questionnaire		
Q	**Question**	**Adapted from**
	Please indicate how much you agree or disagree with the following statements. (1 strongly disagree, 2 moderately disagree, 3 somewhat disagree, 4 neither disagree nor agree, 5 somewhat agree, 6 moderately agree, 7 strongly agree)	
1	Measuring individual signaller workload is important in rail	V
2	Measuring my workload is relevant to my job	V
3	Wearing devices wouldn't require a lot of my mental effort	V
4	A lot of my mental effort would be required to interact with the devices	V
5	I would find the devices difficult to use	V
6	During a shift the devices would be distracting	G, J, P
7	The devices could be a status symbol in my organisation	V
8	I wouldn't use the devices because I would be concerned about being tracked	W
9	Assuming I have access to the devices, I intend to use them	V
10	Given that I would have access to the devices, I predict I would use them	V
11	I would not recommend the devices to my colleagues	New item

References

Gribel, L., Regier, S. and Stengel, I., 2016. Acceptance Factors of Wearable Computing: An Empirical Investigation. In *INC* (pp. 67-72).

Jacobs, J.V., Hettinger, L.J., Huang, Y.H., Jeffries, S., Lesch, M.F., Simmons, L.A., Verma, S.K. and Willetts, J.L., 2019. Employee acceptance of wearable technology in the workplace. *Applied ergonomics*, *78*, pp.148-156.

Parasuraman, A. and Colby, C.L., 2015. An updated and streamlined technology readiness index: TRI 2.0. *Journal of service research*, *18*(1), pp.59-74.

Urquhart, L. and Craigon, P., 2020. The Moral-IT Deck: A Tool for Ethics by Design. *arXiv preprint arXiv:2007.07514*.

Venkatesh, V. and Davis, F.D., 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, *46*(2), pp.186-204.

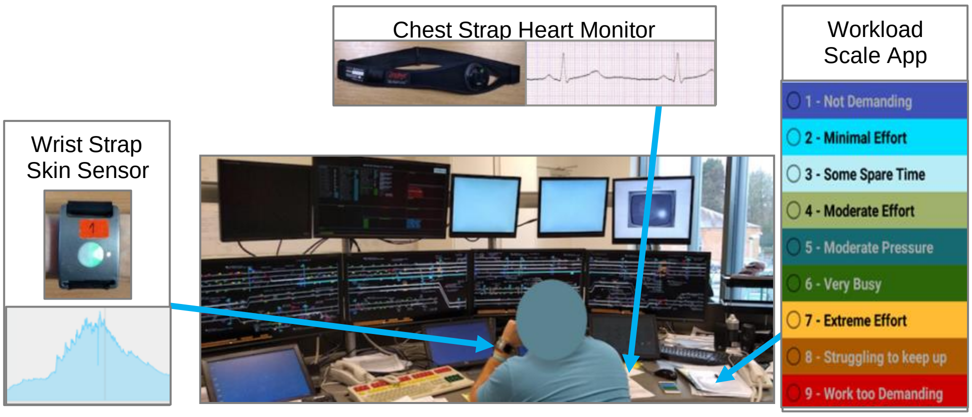
Wolf, P., Menzel, F., Advisory, D.F. and Rennhak, C., 2018. *An Extension of the Technology Acceptance Model tailored to Wearable Device Technology*. Munich Business School Working Paper Series. Abgerufen von https://www. munich-businessschool. de/fileadmin/MBS_Daten/Dateien/Working_Papers/MBS-WP-2018-03. pdf.

Appendix E. Introduction to wearables and app

Attitudes to Wearables in the Workplace

Introduction to Wearables & App

This study assesses attitudes to two potential wearable measures and a self-report app to detect individual workload and effort over a live shift, unobtrusively, and without an observer.



Workload Scale App

- Easy to clean, sits on desk, relies on individual inputs to collect data
- Workload entered every 5 minutes on the mobile device, when prompted by a sound
- Previously used in live signalling operations by giving verbal ratings to an observer
- Also used in simulators by pressing a button on a 1-9 scale
- Deemed not intrusive to the task (7/8 signallers, 1hr simulator trial, NX Panel)
- Shows reported workload. Ratings missed if prompt not heard or concentrating hard

Wrist Strap with Skin Sensor

- Easy to clean, fit, adjust to size, and removeable at any time to stop data collection
- Detects Electrodermal Activity (EDA), temperature, and movement
- Predicted 5 hours comfortable wear time (17/20 participants, 1hr lab-based trial)
- Rarely or never distracting (18/20 participants after 1hr lab-based trial)
- In a signalling study EDA varied depending on phone call content and implications: EDA response was small and recovery quick after a call about a speed restriction; Response was large and recovery slow when a driver said they had hit something.
- In train driving EDA inferred alertness, and anticipation of low adhesion conditions

Chest Strap with Heart Sensor

- Easy to clean, fit, adjust to size, and removeable at any time to stop data collection
- Detects electrical heart activity, breathing rate, movement, posture and location
- Predicted 5 hours comfortable wear time (14/19 participants,1hr lab-based trial)
- Rated rarely or never distracting (17/20 participants after 1hr lab-based trial)
- High workload is inferred from low Heart Rate Variability (HRV) i.e. a steady beat
- In a train driving study HRV decreased at stops and during tunnels
- Note: Recommended not for use by those with a pacemaker (due to Bluetooth)

Use of Data from Wearables

- Data use would be limited to what would be necessary to infer mental workload
- Compliant with GPDR and not to be used to identify someone
- Baseline adjustments allow for factors such as age, gender, hair on wrist etc

Example Uses

- Air Traffic Control researched unobtrusive monitoring in live operations, including how controllers dealt with stress
- Rugby union monitored movement, distance and impact to manage injury risk. Coaches could identify each player's data

How to Use the Measures

Workload Scale App

The screen starts blank

After 5 minutes the scale appears with a short audible alert

You rate how you feel at the time and tap 'finish rating'

The screen returns to blank

IWS is low priority, so you should complete any current tasks before responding

Wrist Strap with Skin Sensor

You wear the skin sensor on your wrist (or on the inside of your wrist)

The light flashes intermittently whilst turned on. It makes no sound.

Data from the device is downloaded to a laptop after data collection.

Chest Strap with Heart Sensor

The chest strap fits around your torso (optional shoulder strap).

Dampening a small area with water improves sensor data collection.

The light flashes intermittently whilst turned on. It makes no sound.

Data from the device is downloaded to a laptop after data collection.

Appendix F. Attitudes study pre-interview questionnaire

Attitudes to Wearables Part 1 v1

Page 1: Consent

This survey aims to collect signalling staff attitudes to wearables for workload measures to help inform how these measures could be used in rail in future. To take part you will be asked to fill in a short questionnaire, be interviewed, then fill in a post-interview questionnaire.

You can contact the researcher Abigail Fowler if you have any questions about the study abigail.fowler@nottingham.ac.uk

Before completing this survey please read the Participant Information Sheet provided by email and note the following:
- Your participation is voluntary. You may withdraw from the study at any time without giving a reason
- Interviews will be audio recorded and transcribed
- All survey data and interview data will be anonymised
- Anonymised results may form part of the PhD Thesis, presentations, academic journal or conference papers
- Anonymised data will be held by the University of Nottingham for a minimum of 10 years, up to 25 years.

You will need to enter a **code provided** to you by the researcher. You will have received this in a text or email.

There are 11 questions.

1. I understand the above information and I agree to participate in this survey * *Required*

○ Yes

2. Date survey completed * *Required*

Dates need to be in the format 'DD/MM/YYYY', for example 27/03/1980.
Please make sure the date is between 14/10/2020 and 31/12/2020.

(dd/mm/yyyy)

3. Enter code provided to you for this study * Required

Your answer should be no more than 5 characters long.

4. Age (in years) * Required

Please enter a whole number (integer).
Please make sure the number is between 15 and 95.
Your answer should be no more than 2 characters long.

5. Gender * Required

- Male
- Female
- Other
- Prefer not to say

6. Date first passed competent as a signaller (If unsure of exact day enter 01, and if unsure of exact month either enter approximate value or 12) * Required

Dates need to be in the format 'DD/MM/YYYY', for example 27/03/1980.
Please make sure the date is between 01/01/1940 and 31/12/2020.

(dd/mm/yyyy)

7. Date first passed competent on a VDU workstation (If unsure of exact day enter 01, and if unsure of exact month either enter approximate value or 12) * Required

Dates need to be in the format 'DD/MM/YYYY', for example 27/03/1980.
Please make sure the date is between 01/01/1940 and 31/12/2020.

(dd/mm/yyyy)

8. Have you ever taken part in a workload assessment at work?

	At **Workstation** * *Required*			In **Simulator** * *Required*		
	Yes	No	If Yes, how many times	Yes	No	If Yes, how many times
You carried out work whilst someone observed for a workload assessment						
You carried out work and rated your workload on a scale 1-9 every few minutes						
You rated workload on 3 scales after you finished your tasks						
You were asked to count the number of items or features on a workstation						
You listed tasks in a typical hour at a new workstation, in a workshop						
Other (please state in next question)						

9.a. Please list wearables or fitness apps you have used

Individual Attitude to Change

This survey uses the Innovativeness Scale (Hurt et al., 1977), designed to measure individuals' orientations towards change.

People respond to their environment in different ways. This questionnaire measures an individual's orientations toward change in your own life. The statements refer to some of the ways people can respond. Please indicate the degree to which each statement applies to you.

1 Strongly Disagree, 2 Disagree, 3 Neutral, 4 Agree, 5 Strongly Agree

Please work quickly, there are no right or wrong answers, just record your first impression.

(20 statements in total)

	1 Strongly Disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly Agree
My peers often ask me for advice or information	○	○	○	○	○
I enjoy trying new ideas	○	○	○	○	○
I seek out new ways to do things	○	○	○	○	○
I am generally cautious about accepting new ideas	○	○	○	○	○
I frequently improvise methods for solving a problem when an answer is not apparent	○	○	○	○	○
I am suspicious of new inventions and new ways of thinking	○	○	○	○	○
I rarely trust new ideas until I can see whether the vast majority of people around me accept them	○	○	○	○	○
I feel that I am an influential member of my peer group	○	○	○	○	○
I consider myself to be creative and original in my thinking and behaviour	○	○	○	○	○
I am aware that I am usually one of the last people in my group to accept something new	○	○	○	○	○

	1 Strongly Disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly Agree
I am an inventive kind of person	○	○	○	○	○
I enjoy taking part in the leadership responsibilities of the group I belong to	○	○	○	○	○
I am reluctant about adopting new ways of doing things until I see them working for people around me	○	○	○	○	○
I find it stimulating to be original in my thinking and behaviour	○	○	○	○	○
I tend to feel that the old way of living and doing things is the best way	○	○	○	○	○
I am challenged by ambiguities and unsolved problems	○	○	○	○	○
I must see other people using new innovations before I will consider them	○	○	○	○	○
I am receptive to new ideas	○	○	○	○	○
I am challenged by unanswered questions	○	○	○	○	○
I often find myself sceptical to new ideas	○	○	○	○	○

Finished

Thank you for participating in this survey.

Scoring:

Step 1: Add the scores for items 4, 6, 7, 10, 13, 15, 17, and 20.

Step 2: Add the scores for items 1, 2, 3, 5, 8, 9, 11, 12, 14, 16, 18, and 19.

Step 3: Complete the following formula: II = 42 + total score for Step 2 - total score for Step 1.

Scores above 80 are classified as Innovators.

Scores between 69 and 80 are classified as Early Adopters.

Scores between 57 and 68 are classified as Early Majority.

Scores between 46 and 56 are classified as Late Majority.

Scores below 46 are classified as Laggards/Traditionalists.

In general people who score above 68 and considered highly innovative, and people who score below 64 are considered low in innovativeness.

Hurt, H. T., Joseph, K., and Cook, C. D. (1977). Scales for the measurement of innovativeness. Human Communication Research, 4, 58-65

Appendix G. Attitudes study interview questions

	Topic	Interview Question
1	Experience	If you use wearables, what's your favourite and why?
2.1	**Devices** Perceived Ease of Use	How distracting could devices be for you during a shift? (chest, wrist, app)
2.2		How comfortable could devices be for you during a shift? (chest, wrist, app) *(*NOTE re COVID 19 - straps are washable)
3	Subjective Norm/ Image	What is the likely reaction of colleagues to the use of these new measures (chest, wrist, app)?
4	Own experiences of workload	From your experience on shift give examples of: • Anticipating a problem correctly (a 'gut' reaction) • Time pressure/ stress • Unexpected event/ system response not expected • Brief workload peak affected you for rest of shift • Confidence when successful achieving your tasks • Impact of a change of new technology or procedures

Use Cases

D.	**Use Case** Understand Signaller Workload	Assess *task* – peaks/troughs in workload & overall effort Devices Infer – anticipation, alertness, stress, time pressure, brief peaks in reaction to unexpected events etc Data – anonymised When – 1-2 shifts
5.1	Perceived Usefulness	How useful to demonstrate to *others how hard you work*?
5.2	Anonymity	Acceptable?
5.3	Job Relevance	What's inferred – which most relevant to assessing *task*?
5.4	Trust	Any concerns about data use?
5.5	Time	Data collection needed vs tolerable: >4 hrs? Continuous or every 5 mins (app)
E.	Learning Aid	Assess *trainee* – track progress, self-learning & training effectiveness. Infer – alertness, confidence, unexpected events, stress, effort Data – You & trainer share. (Device supplier?) When – Before/during/after training.
6.1	Perceived Usefulness	How useful for *you* to understand your data (e.g. heart)?
6.2	Anonymity	Acceptable?
6.3	Job Relevance	What's inferred – which most relevant to assessing *trainees?*
6.4	Trust	Any concerns about data use?
6.5	Time	Data collection needed vs tolerable: >4 hrs? Continuous or every 5 mins (app)
F.	Assess impact of new tech or procedures	Assess *change* – work support/ effectiveness of change Infer – stress, effort, unexpected system responses Data – Workstation & initials. Managers may know. Supplier? Investigator? When – Before/during/after change.
7.1	Perceived Usefulness	How useful to demonstrate to *others impact of changes?*
7.2	Anonymity	Acceptable?
7.3	Job Relevance	What's inferred – which most relevant to assessing impact of *change*?
7.4	Trust	Any concerns about data use?
7.5	Time	Data collection needed vs tolerable: >4 hrs? Continuous or every 5 mins (app)

8	Precautionary Principle	Just because we could use these measures, should we?

Appendix H. Attitudes study post-interview questionnaire

Attitudes to Wearables Post Interview Questionnaire v1

Page 1: Post-Interview Questionnaire

This **Post-Interview Questionnaire** collects signalling staff attitudes to wearables for workload measures to help inform how these measures could be used in rail in future.

You can contact the researcher Abigail Fowler if you have any questions about the study abigail.fowler@nottingham.ac.uk

Before completing this survey please read the Participant Information Sheet provided by email and note the following:
- Your participation is voluntary. You may withdraw from the study at any time without giving a reason
- Interviews will be audio recorded and transcribed
- All survey data and interview data will be anonymised
- Anonymised results may form part of the PhD Thesis, presentations, academic journal or conference papers
- Anonymised data will be held by the University of Nottingham for a minimum of 10 years, up to 25 years.

You will need to **enter a code** provided to you by the researcher. You will have received this in a text or email.

There are 12 questions.

1. Please enter the code provided to you for this study (letter and number code) *
Required

Your answer should be no more than 5 characters long.

2. Please indicate how much you agree or disagree with the following statements. Please answer quickly, there is no right or wrong answer

	* Required						
	1 strongly disagree	2 moderately disagree	3 somewhat disagree	4 neither disagree nor agree	5 somewhat agree	6 moderately agree	7 strongly agree
Measuring individual signaller workload is important in rail	○	○	○	○	○	○	○
Measuring my workload is relevant to my job	○	○	○	○	○	○	○

The rest of this survey covers 3 potential future devices to collect data in workload assessments:

- Chest strap heart monitor
- Wrist strap skin sensor
- App on a phone

Check if statements say you WOULD or WOULDN'T.

Please indicate how much you agree or disagree with the following statements. Please answer quickly, there is no right or wrong answer.

3. Wearing devices wouldn't require a lot of my mental effort

	* Required						
	1 strongly disagree	2 moderately disagree	3 somewhat disagree	4 neither disagree nor agree	5 somewhat agree	6 moderately agree	7 strongly agree
Chest	○	○	○	○	○	○	○
Wrist	○	○	○	○	○	○	○

4. A lot of my mental effort would be required to interact with the devices

	* Required						
	1 strongly disagree	2 moderately disagree	3 somewhat disagree	4 neither disagree nor agree	5 somewhat agree	6 moderately agree	7 strongly agree
Chest	○	○	○	○	○	○	○
Wrist	○	○	○	○	○	○	○
App	○	○	○	○	○	○	○

5. I would find the devices difficult to use

	* Required						
	1 strongly disagree	2 moderately disagree	3 somewhat disagree	4 neither disagree nor agree	5 somewhat agree	6 moderately agree	7 strongly agree
Chest	○	○	○	○	○	○	○
Wrist	○	○	○	○	○	○	○
App	○	○	○	○	○	○	○

6. During a shift the devices would be distracting

	* Required						
	1 strongly disagree	2 moderately disagree	3 somewhat disagree	4 neither disagree nor agree	5 somewhat agree	6 moderately agree	7 strongly agree
Chest	○	○	○	○	○	○	○
Wrist	○	○	○	○	○	○	○
App	○	○	○	○	○	○	○

7. The devices could be a status symbol in my organisation

	* Required						
	1 strongly disagree	2 moderately disagree	3 somewhat disagree	4 neither disagree nor agree	5 somewhat agree	6 moderately agree	7 strongly agree
Chest	○	○	○	○	○	○	○
Wrist	○	○	○	○	○	○	○
App	○	○	○	○	○	○	○

8. I wouldn't use the devices because I would be concerned about being tracked

	* Required						
	1 strongly disagree	2 moderately disagree	3 somewhat disagree	4 neither disagree nor agree	5 somewhat agree	6 moderately agree	7 strongly agree
Chest	○	○	○	○	○	○	○
Wrist	○	○	○	○	○	○	○
App	○	○	○	○	○	○	○

9. Assuming I have access to the devices, I intend to use them

	* Required						
	1 strongly disagree	2 moderately disagree	3 somewhat disagree	4 neither disagree nor agree	5 somewhat agree	6 moderately agree	7 strongly agree
Chest	○	○	○	○	○	○	○
Wrist	○	○	○	○	○	○	○
App	○	○	○	○	○	○	○

10. Given that I would have access to the devices, I predict I would use them

	* Required						
	1 strongly disagree	2 moderately disagree	3 somewhat disagree	4 neither disagree nor agree	5 somewhat agree	6 moderately agree	7 strongly agree
Chest	○	○	○	○	○	○	○
Wrist	○	○	○	○	○	○	○
App	○	○	○	○	○	○	○

11. I would not recommend the devices to my colleagues

	* Required						
	1 strongly disagree	2 moderately disagree	3 somewhat disagree	4 neither disagree nor agree	5 somewhat agree	6 moderately agree	7 strongly agree
Chest	○	○	○	○	○	○	○
Wrist	○	○	○	○	○	○	○
App	○	○	○	○	○	○	○

12. Please add any other comments you have relevant to this study

Finished

Thank you for participating in this survey.