Comparison of real-time relative workload measurements in rail signallers*

By

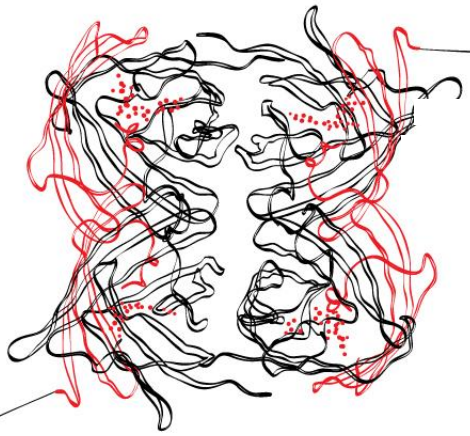
Rob van Broekhoven

Supervised by:

1. Willy Siegel, M.Sc.
2. Prof. Dr. Jan Maarten Schraagen

University of Twente, Human Factors & Engineering Psychology
Master of Science Program
M.Sc. Thesis

22th March 2016



*This master thesis is an expanded version of a conference paper published in the Proceedings of the German Workshop on Rail Human Factors, held in Braunschweig on the 8th and 9th of March 2016: Broekhoven, R., Siegel, A. W., Schraagen, J. M., & Noordzij, M. L. (in production). Comparison of real-time relative workload measurements in rail signallers.

UNIVERSITY OF TWENTE.

Samenvatting

Deze exploratieve veldstudie richt zich op zwakke veerkrachtsignalen van werklast in een verkeerspost bij treindienstleiders. Dit onderzoek richt zich op de vraag of een ondersteunend real time systeem informatie werklast van treindienstleiders kan voorspellen (Siegel & Schraagen, 2014) en of deze methode verbeterd kan worden. Om deze vraag te beantwoorden werden drie maten gebruikt om werklast te meten. De eerste was een subjectieve geïntegreerde werklastschaal. De tweede was een fysiologisch meetinstrument dat gebruik maakte van huidgeleiding. De derde maat was het systematisch observeren van gedrag van treindienstleiders. De systeeminformatie, huidgeleiding en gedragsobservaties werden vergeleken met de subjectieve maat voor werklast voor drie casussen. De resultaten laten zien dat de systeeminformatie over communicatie, handmatige acties en het schakelen tussen taken onderscheidend is voor het voorspellen van de subjectieve werklast. De systeem informatie over monitoren en planning waren daarentegen niet discriminerend voor het voorspellen van subjectieve werklast. Met informatie over de eigen werklast zouden treindienstleiders meer inzicht kunnen krijgen in het eigen functioneren en zo de veerkracht van de post vergroten.

Abstract

This exploratory field study investigated the weak resilience signals of workload in a rail traffic control room. The goals of this research are to see whether real-time system information of a rail control post can be used to predict workload of a rail signaller in real-time (Siegel & Schraagen, 2014) and to further improve this method. In order to investigate this question, three workload measures were used. The first was the subjective Integrated Workload Scale, the second was a physiological measurement of electrodermal activity and the third was behavioural observation. For three cases, the subjective workload was compared to the system information algorithm and the two other workload measures. The results show that the system information of communication, manual actions and switch cost are discriminating for workload. The system information about monitoring and planning did not seem to discriminate between different levels of workload. This study validates that system information can be used to predict workload. With this information insight in functioning of the rail signaller could be enhanced, which can improve the resilience of the rail post.

Keywords: workload, system data, IWS, electrodermal activity, rail control room, rail signaller

Introduction

Resilience in rail socio-technical systems

Society and complex systems get ever more complex and the regulation of these complex systems needs follow stricter safety norms (Amalberti, 2001). Socio-technical systems consist of interactions between human and complex technical systems. The rail control post and the rail signallers play a central role in the rail socio-technical system (Belmonte, Schon, Heurley, & Capel, 2011). Rail signallers have an overview of trains from different organisations and monitor interactions between the rail system and other infrastructural networks such as waterways and road traffic in case these networks cross a railway. These rail signallers monitor the rail system for calamities, delays and disturbances, and regulate the traffic when problems arise. In this study, the functioning of rail signallers is approached from a resilience engineering perspective. Resilience engineering focuses on how socio-technical systems handle unexpected and unforeseen circumstances (Hollnagel, Woods, & Leveson, 2006). One representation of how such a system can be described, is the Dynamic Safety Model of Rasmussen (1997; Figure 1). This model states that for a system to operate in a normal way, it should remain between three boundaries: the performance or safety boundary, the economic boundary and the workload boundary. The model describes that the system is under constant influence to uphold safety procedures, to be cost efficient and to get the work done with the least effort possible.

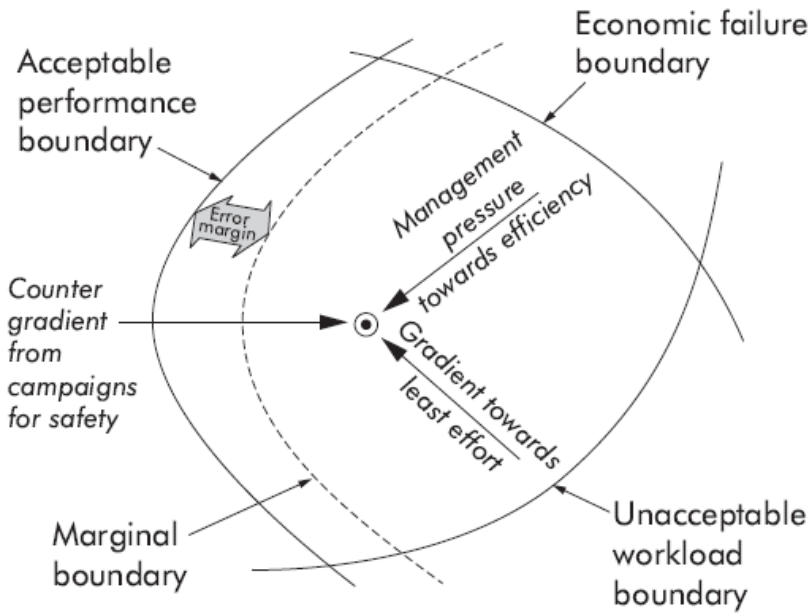


Figure 1. modified dynamic safety model from Rasmussen (1997), copied from Cook & Rasmussen (2005).

These boundaries, however, are not predefined and they are different for every system and every situation. It is therefore difficult to determine the real boundary for a particular socio-technical system. Accidently crossing a boundary and risking a dangerous situation, could happen without being aware of this. If no accidents happen for a longer period of time, the impression might be that no boundary is crossed and that the efficiency is increased or workload is decreased, and causing the organisation to move even further or longer over the boundary (Cook & Rasmussen, 2005). So crossing a boundary is not always immediately visible, nor will it immediately result in accidents. It will, however, reduce the extra resources needed to handle unexpected situations. This might lead to system failures when unforeseen or abnormal circumstances arise (Woods, Wreathall, & Anders, 2006), which might eventually result in accidents, overburdened personnel or financial problems.

Siegel and Schraagen (2014) proposed a method that could help to prevent crossing these boundaries and to have some indication whether a boundary is approaching or not. To accomplish this, the socio-technical systems should focus on so-called 'weak resilience signals'. Weak resilience signals are signals that indicate a possible degradation of the socio-technical system without immediately triggering a predefined alarm. These signals can help to prevent crossing a boundary inadvertently. An example of a weak resilience signal could be a change in experienced workload that is not noticed or is not recognized as an alarming signal. For an organisation, both Madni and Jackson (2009) and Hollnagel (2009) state that the level of resilience is not merely a given factor, but an ability that can be developed to make the organization more flexible and proactive, to be better able to handle unforeseen events. Important factors in developing resilience are the ability and opportunity to anticipate, monitor, respond and learn from situations (Hollnagel, 2009).

To facilitate learning, retrieving good and accurate information is important. One way of retrieving information could be by processing and analysing relevant system information from the socio-technical system. This is for example done by Zeitlin (1998) in automotive industries. He used system data such as traffic density, speed and brake application to estimate mental workload in van drivers. Ohm and Ludwig (2013) also used acceleration and speed detection retrieved from a mobile phone in the car and found that the cognitive load of the driver could be predicted with 76% accuracy with this simple data. In automotive studies it seems that with relative simple system information, a good indication of cognitive workload can be given.

To improve resilience in a professional rail setting, Siegel and Schraagen (2014) developed a real time support system utilizing system data retrieved from rail operator ProRail. One of the investigated weak resilience signals described by Siegel and Schraagen (2014) is the relative increase or decrease of subjective and objective workload. Presenting these weak resilience signals was done by showing rail signallers with changes of their subjective workload and objective workload calculated

from their traffic information system. Subjective workload was operationalized by a one-dimensional workload scale designed for rail signallers. This scale is called the Integrated Workload Scale (IWS; Pickup, Wilson, Norris, Mitchell, & Morrisroe, 2005). Objective workload was operationalized by means of an algorithm based on the model of cognitive task load (CTL; Neerincx, 2003). The cognitive task load for a certain task is based on three dimensions: task complexity, task duration and task switching. The more complex, the longer the duration or the more switching between different tasks, the higher the cognitive task load. Siegel and Schraagen (2014) developed an algorithm, derived from system information and log data of the traffic system, resulting in a measure called the external cognitive task load (XTL; Siegel & Schraagen, 2014). The XTL is based upon four main measurable tasks of the rail signaller: monitoring, plan mutations, manual actions and communication by telephone. Because these measures are taken from system information, there could be a discrepancy between the behaviours that the system data would predict and the actual behaviour of the rail signaller. For example, an automatic mutation in the planning can change into something else without the rail signaller's awareness. This will have no effect on the executed behaviours of the rail signaller, but the system will record the activity. This study will therefore investigate whether socio-technical system data in the XTL algorithm are a good and valid method in predicting workload in professional rail signallers, by comparing the XTL algorithm to other workload measures.

Workload measures

Workload has been extensively investigated but there is no consistent definition of workload in the literature (e.g. Young, Brookhuis, Wickens, & Hancock, 2015). The problem is that there is no exact empirical definition and no physical unit to measure workload. Generally, the literature suggests that a good analogy of workload consists of two components: the "stress side" composed of task demands or task load and the "strain side" that stands for the impact on a human being (Schlegel, 1993). This stress/strain comparison offers the opportunity to look at workload as a multidisciplinary construct. On the one hand, it has multiple facets on the stress side, like time pressure and task complexity. On the other hand, it also gives the opportunity to look at different aspects on the strain side, such as how workload is experienced and available cognitive capacity. Resulting from these different perspectives on workload a whole range of methods attempt to measure workload. These methods generally focus on different facets of workload, such as self-report questionnaires (NASA-TLX; Hart & Staveland, 1988), heart rate variability (Jorna, 1992) involuntary eye movement (Obinata, Tokuda, Fukuda & Hamada, 2009) and EEG (Brookhuis & de Waard, 2011). There is consensus, however, that at least

three components are important for measuring workload. These components are subjective, physiological and performance measures (Young et al., 2015). Therefore, the current study will use three different ways of measuring workload, corresponding to these three components.

First of all, for the subjective measure it was important that the rail signallers were not burdened more than necessary, since the rail signallers are working in a real situation during measurement. Questionnaires like the NASA-TLX (Hart & Staveland, 1988), the SWAT (Reid & Nygren, 1988) and the WP (Tsang & Velazquez, 1996) were not an option because they take over 60 minutes to administer (Rubio, Diaz, Martin, & Puente, 2004). Therefore we used the Integrated Workload Scale (IWS; Pickup et al., 2005). The IWS consists of a 9-point one-dimensional scale on which the rail signallers can indicate their perceived workload for a certain period of time. The IWS is specifically designed and implemented to measure a rail signaller's subjective workload and it gives an insight in their perceived cognitive workload while being easy to fill in and not being intrusive on the work of the rail signaller.

Second, for the physiological measure, we used electrodermal activity (EDA). Electrodermal activity is an ongoing physiological measure of workload and there is consensus that it at least reflects a general measure of arousal or stress (Healey & Picard, 2005). The EDA is expressed in skin conductance (SC) units (Boucsein, 2012). In the EDA measurement, there are several parameters that can be extracted. Some parameters are related to the phasic, short lived Skin Conductance Responses (SCR), whereas others are related to tonic, slow changes in the average level of the skin conductance level (SCL). The EDA measurement directly reflects activity of the sympathetic nervous system without being affected by parasympathetic activity (Boucsein, 2012) and it is a non-intrusive measurement that minimizes motion artefacts (Poh, Swenson, & Picard, 2010). As the EDA measurement is not intrusive, the rail signallers will not be disrupted in their work.

Third, for the performance measure we used state sequence based behavioural observation (Bakeman, 2000). In state sequence coding, the behaviours are a string of coded events or states that preserves time information (duration) and can be compared to the measures obtained from the XTL (Siegel & Schraagen, 2014). The behavioural observation will focus on the executed behaviours of the rail signaller and will form the basis for a behavioural performance measure. We expect that certain behaviours correlate with the XTL measure, but that some differentiations in behaviours go undetected by the XTL.

The current research builds upon a previous study by Siegel and Schraagen (2014) by adding behavioural observations and EDA in measuring workload in comparison to the on system based data XLT algorithm. Furthermore, the other measurements can be used to further calibrate this algorithm. This exploratory field study attempted to answer two research questions. The first is whether the four

workload measurements (IWS, XTL, EDA and behavioural observations) support each other in the identification of changes in observed workload. The second question is whether the objective workload measure XTL employed by Siegel and Schraagen (2014) can be compared and complemented with these measurements to better match workload variations.

Methods

Research setting & Participants

The observations took place in a rail control post in Alkmaar. In this rail control post, there are four rail signallers on active duty, one backup rail signaller for calamities and to fill in breaks, one decentralized train traffic manager and one team supervisor. On the post present were also one bridge controller for waterways in and around Alkmaar and up to three public transport servants that communicated changes and delays to the public (Figure 2). This observational study will focus on one of the four workstations (WS) of the rail control post where a rail signaller will be observed. The workstation most observed was workstation 2 for the region Uitgeest. This workstation was indicated to have the most workload and had a high influx of trains during rush-hour, and would be the best workstation to measure different levels of workload.

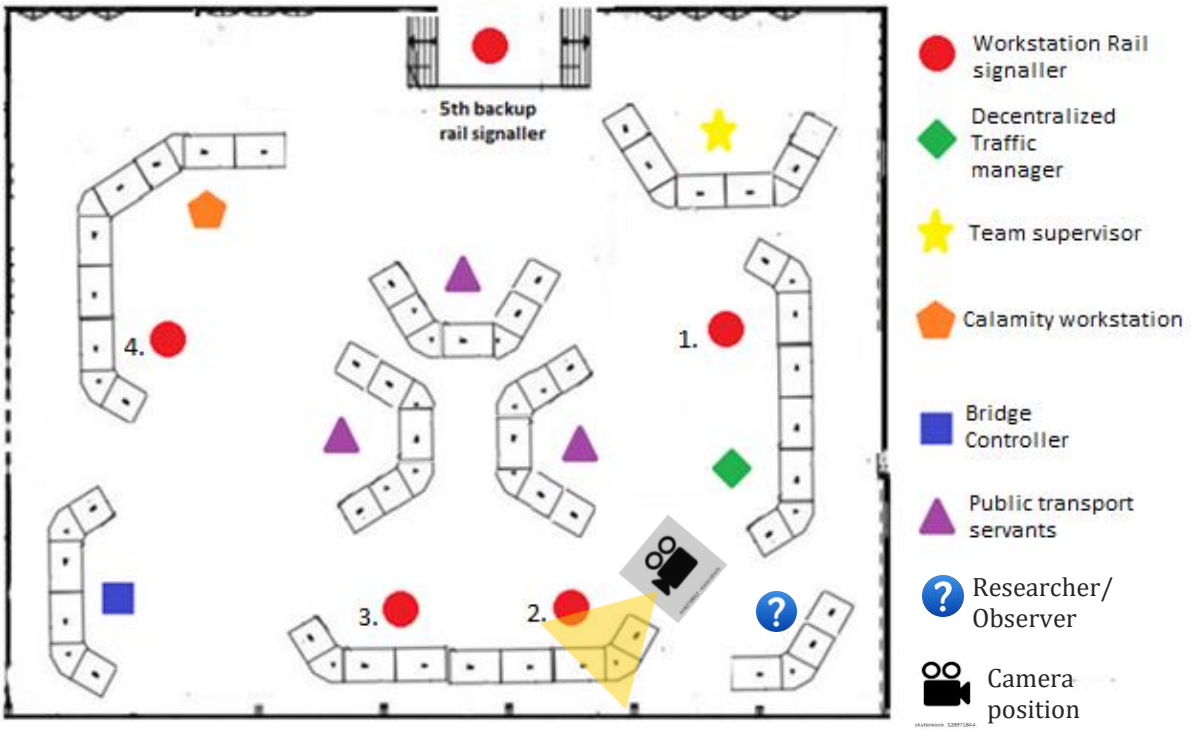


Figure 2. Rail control post Alkmaar layout.

The work of a rail signaller consists for large parts of the shift of monitoring the situation, and adjusting the automatic systems if the prescheduled plan cannot be executed automatically. This is for example the case when a train is delayed. Besides the tasks of monitoring and planning, the rail signaller also has to intervene when an incident occurs and a possible dangerous situation arises. If this occurs, the rail signaller must take manual control of the situation. At that moment, the rail signaller and colleagues must execute a lot of different actions, such as alerting direct colleagues and other rail control posts and calling train drivers while maintaining an overview of the situation. They have to maintain this overview to reroute or stop incoming trains to avoid accidents. An example of such an incident, and the main case of this paper will give a short illustration of the work of a rail signaller (Table 1),

Table 1

Description of rail signaller word, case 1

It is Wednesday 14:00 h. at a rail control room in Alkmaar, The Netherlands. A rail signaller throws a glance at his screen and makes some adjustments in the rail traffic planning. It is a calm shift and the train traffic runs normally. Time passes by and at 16:23, a train driver calls in to report that the train has hit an object. The driver has stopped the train to check out what happened, as the procedure prescribes. As a response to the incoming call, the rail signaller notifies the decentralized traffic manager and rail signaller of the adjacent area of the stop location of the train. Next, he proceeds to inform the rail signaller of another adjacent post about the situation. Because the situation is rather unclear, the rail signaller calls all the approaching trains and orders them to stop. Each call includes exact prescribed actions and mileage to avoid miscommunication. After seven minutes (16:30) the inspecting rail driver reports that he did not find anything that could explain the sound he heard and that there is no sign that the train has hit a person. Therefore, the rail signaller gives permission to drive again. The local co-workers, the decentralised traffic manager and the rail signaller from the other post are informed and the restrictions are cancelled. The rail signaller calls all related trains to abrogate the restrictions and informs them that they may start driving again. He requests to remain vigilant around the reported area. This case presents the effects of one train stopping for 7 minutes with the consequences of a workload increase for more than half an hour. Around 17:00, the last telephone call was conducted. While the events unfolded, the rail signaller had to monitor and act on different trains and events. The rail signaller was constantly switching between incoming calls from train drivers, informing co-workers, being updated by co-workers, anticipating on all new incoming trains in the area, manually rerouting these trains and informing all involved train drivers by telephone. This case describes a possible urgent and alarming situation where lots of different actions are necessary and a lot of different people need to communicate. The rail signaller had to make a fast switch between a low task demand and critical situations where multiple actions must be handled simultaneous.

The 10 rail signallers (9 male and 1 female) that participated in this study were between 22 and 52 years old (M = 37.6; SD = 11.12) and had experience ranging from half a year up to 34 years (M = 11.8; SD = 11.01). Participants participated voluntarily and did not receive any financial compensation.

Measurements

Integrated Workload Scale (IWS)

Siegel and Schraagen (2014; Figure 3) developed a tool based on the Dutch validated IWS (Wilms & Zeilstra, 2013) that is a translation of the English IWS scale of Pickup and colleagues (2015). In the current study the tool developed by Siegel and Schraagen (2014) was used. The IWS tool would pop up every five minutes for a duration of 30 seconds on the rail signaller's work station. In this way, we received an automatic and continuous rating of the IWS tool. To maintain a high response rate, it was possible for the rail signallers to open and fill in the IWS tool during the whole five minutes. It was also possible to adjust the value of the previous five minutes. This gave the rail signaller the opportunity to primarily focus on handling the situation, while still having the ability to fill out the IWS tool. If no response was given, the last value was copied under the assumption that there was no change of experienced workload. A longer duration of a high IWS will be referred to as a "stretch" of increased subjective workload as defined in Siegel and Schraagen (2014). In order to compare the methods with each other, we took the IWS pattern retrieved from the IWS tool as a baseline to make a distinction between low (IWS, 1-2) and high (IWS, 3-9) workload. We chose for IWS as a baseline because it has a uni-dimensional scale and because the IWS is used before in a rail setting and is validated for rail controllers (Pickup et al., 2005; Wilms & Zeilstra, 2013). The IWS measure therefore provides a good and easy baseline measure.

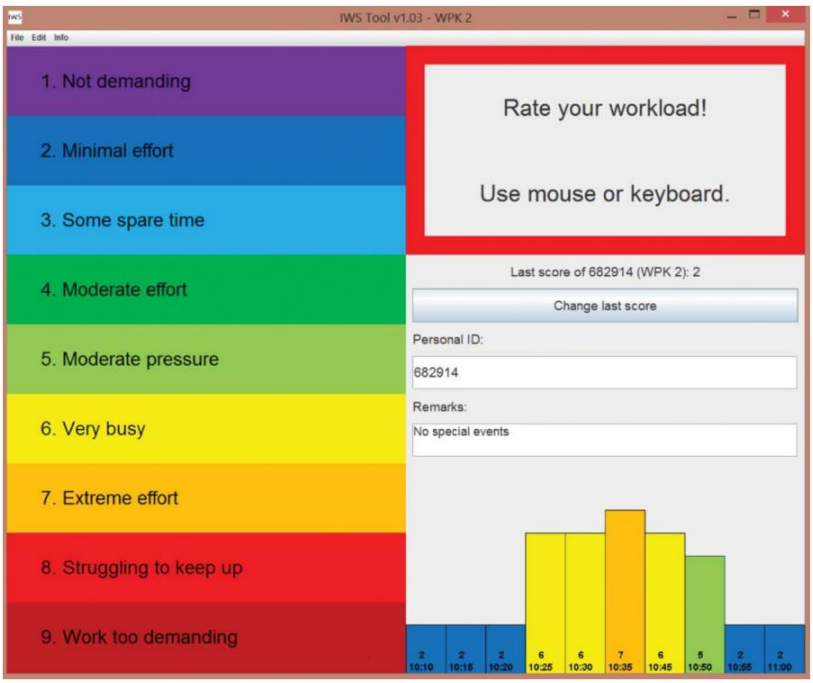


Figure 3. Integrated Workload Scale tool (Siegel & Schraagen, 2014), translated Dutch to English.

Electrodermal activity (EDA)

Rail signallers were asked to wear the Affectiva QTM sensor. This is a wrist worn, watch-like sensor that measures EDA with 1 cm diameter Ag-AgCl dry electrodes at the ventral side of the wrist. EDA data were pre-processed with a Continuous Decomposition Analysis (CDA) as implemented in Ledalab (Benedek & Kaernbach, 2010), which requires MATLAB (Mathworks, Natick, MA, USA). From the EDA, an estimate of the skin conductance level (SCL) as well as the overlaying phasic activity (occurrence and amplitude of SCRs) can be acquired. The phasic activity, coming from classical Trough-to-Peak analysis, was reported (threshold for an SCR amplitude was set at .03 µS; Boucsein, 2012). As recommended by Boucsein (2012), visual checks were performed on plots of skin conductance data to identify failed measurements, "non-responding" (indicated by an absence of SCRs in a given measurement) and incorrect classification of SCRs. Data from these problematic measurements were removed from further analysis. The SCL and SCR parameters were expressed in 5 minute intervals to allow for comparisons to the XTL and the input in the IWS tool.

External Cognitive Task Load (XTL)

The External Cognitive Task Load (XTL) was calculated from real-time data retrieved from the operational control system (Siegel & Schraagen, 2014). The XTL was adjusted by adding a one to the formula used by Siegel and Schraagen (2014). This was done to achieve the same range as the IWS tool (from 1 to 9). The algorithm was based on the number of automatically executed plan rules in 5 minutes per workstation (monitoring, *mon*), the number of mutated plan rules in 5 minutes per workstation (plan mutations, *plan*), the number of non-executed plan rules in 5 minutes per workstation (manual actions, *man*), and the percentage of seconds spoken through the telephone of 5 minutes per workstation (communications, *com*). The constant k's were based on the adjusted k's from the study of Siegel and Schraagen (2014) that optimized the relation with IWS tool: K_{mon} = 0.4, $K_{plan} = 0.9$, K_{man} = 1.2 and $K_{com} = 1.5$.

$$XTL_{WS} = K_{switchWS}\left(K_{mon}\frac{Mon_{WS}}{Mon_{max}} + K_{plan}\frac{Plan_{WS}}{Plan_{max}} + K_{man}\frac{Man_{WS}}{Man_{max}} + K_{com}\cdot Com_{WS}\right) + 1 \quad (1)$$

$$1 \leq XTL_{WS} \leq 9 \; ; \sum_{i=1}^{4} K_i = 4$$

The XTL formula of Siegel and Schraagen (2014) has been altered by adding a 1, causing the XTL values to be between 1 and 9, just like the values retrieved from the IWS tool.

The switch cost was taken into account by the number of activations that is composed of: 1) the number

of delayed trains, 2) the number of telephone calls, and 3) the number of incidents reported in 5 minutes per workstation divided by the maximum number of activations in a 5 min time slot. The XTL gives a general relative cognitive task load configured from system output each five minutes. It will also provide a relative load of each of the four categories (monitoring, planning, manual and communication) which can be used to look at specific components in the XTL formula.

$$K_{switchWS} = \frac{number\ of\ activations\ in\ 5\ \text{min}\ baseslot}{maximum\ number\ of\ activations\ in\ 5\ \text{min}\ baseslot} + 1 \qquad (2)$$

$$1 \leq K_{Switch} \leq 2$$

K_{switch} formula as used in Siegel & Schraagen (2014)

Behavioural cognitive Task Load (BTL)

The Behavioural cognitive Task Load (BTL) was calculated in a similar way as the XTL. The BTL is based on the model of Neerincx (2003) to be able to compare the variables of both measures with each other. The difference between the XTL and the BTL is that the information for the XTL comes from the ProRail information system, whereas the information from the BTL comes from behaviours of the rail signaller. Behaviours were selected based on interviews with rail signallers, observations and the four different categories of the XTL. Behaviours were recorded with a webcam and processed in the software program Observer XT (version 11). This was done based on how long (s) a behaviour was executed and how many switches occurred between different behaviours in five-minute time frames. This was done by observing for how long rail signallers showed behaviours that were linked to monitoring, manual actions, planning behaviour, communication with team members and making telephone calls with others outside the rail traffic control post. These categories were further specified, taking into account different behaviours and implementation locations (Table 2; Figure 4). A differentiation was made, for example, between telephone calls originating from different parts of the socio-technical system. More specifically, a call from a bridge operator is likely to cause a low increased workload because the waterway bridge is manually controlled with one button. On the other hand, an incoming alarm call is more likely to increase workload because it needs immediate action.

Table 2

Overview categorized behaviours BTL

Monitor	Planning	Manual	Communication	
Fast and global glance on the screens	Manual plan screen	Rail occupancy screen	Local communication	Communication trough telephone
Rail occupancy screen	Plan screen monitor	Overview screen	Decentralized traffic regulator	Bridge
Overview screen		Writing report	Co-RS, specific case	Co-RS-in other post
		Other	Co-RS or other co-worker, general but work related	Train driver
				Alarm

Figure 4. Screen one is the rail occupancy screen, screen two is the planning screen, the three screens under number three are the overview screens

The formula of the BTL is based on the time(s) that behaviours in the categories (mon, plan, man, com) were observed. The constant k's were initialized with 1.

$$BTL_{ws} = B_{switchWS}\left(k\,mon * mon\,(s) + k\,plan * plan\,(s) + k\,man * man\,(s) + k\,com * com\,(s)\right) \quad (3)$$

Again, the factor 'switch cost' from Neerincx (2003) was integrated. The switch cost for BTL [4] was based on the number of switches in 5 minute intervals divided by the maximum observed number of changes in behaviour. The maximum number of behavioural switches observed during the study was 60 in 5 minutes.

$$B_{switchWS} = \frac{number\ of\ different\ behaviors\ in\ five\ minutes\ slot}{maximum\ number\ of\ different\ behaviors\ in\ a\ five\ minute\ slot} + 1 \quad (4)$$

Procedure

The protocol guiding the observations included an oral recorded consent instead of a written consent. This was due to cultural constraints and the specific request of the post management, and was approved by the ethics committee of the University of Twente (registration number: BCE15412). Before the observations started, the instructions and goals of the study were explained. When the participant was ready and everything was clear, the participant was asked to wear the EDA-sensor and was informed that the behavioural observations would start in a few minutes. The IWS tool measurement was running during the whole day and evening. The EDA measurement, as well as camera monitoring, was conducted during the day provided that participants were willing to wear the EDA sensor and agreed to be recorded. In total, 34 hours of EDA measurement and 26 hours of behavioural observations were recorded. The different observations had a duration ranging from one hour up to three hours before the end of the shift or a break. The camera monitoring was done to capture possible unique events and to look back for specific behaviours. The camera was positioned to offer a view from the side; this setup gave the opportunity to observe the behaviour of the rail signaller but also to see and confirm system mutations displayed on the screen as well as to confirm manual actions on the different screens (Figure 5).

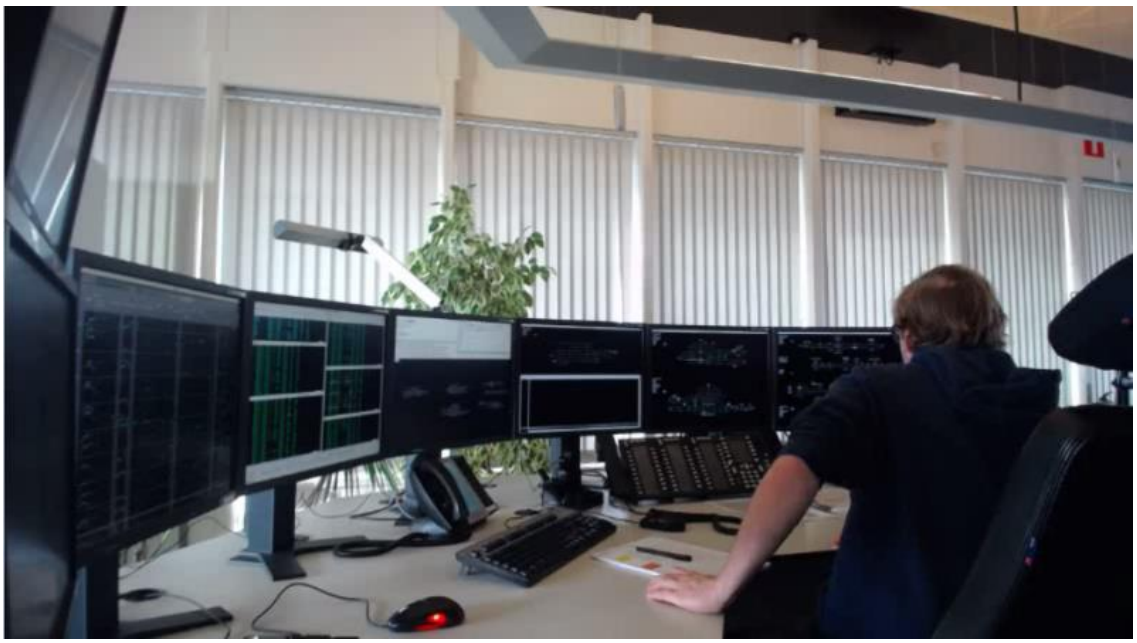


Figure 5. View of the camera position used, showing both the behaviour of the rail signaller and changes on the rail traffic screens.

Coding of observed behaviour was restricted to half an hour before the subjective IWS tool measure indicated “Some spare time” or higher. This was done for practical reasons in analysing all recorded material. During and between shifts, it was possible for the rail signallers to rotate positions. If this happened, the EDA-sensor was retrieved and data were extracted and logged before the EDA-sensor was passed on to the next rail signaller. Camera and behavioural observations continued, but a change of shift was marked in the video file. When the shift was coming to an end, the participants were asked for any remarks about the shift and were thanked for their participation.

Results

Data collection and case comparison

In order to compare the methods with each other, the IWS tool data was used as a baseline measure. However, occurrences of incidents of high workload were rare during the study. Therefore, behavioural observation was further analysed only around IWS elevations. During our observations, the IWS rose 14 times above “minimal effort (2)” and once a longer period of “very busy”(6)/extremely busy”(7) was indicated. In three of these elevations, the IWS pattern showed a clear stretch in the IWS and the data collected from the other measurements were usable. This provided us with 3 cases (Table 3). Two of these cases (Cases 1 and 2) contained sufficient data points for further statistical analyses. In case 3, the sample size of data points was smaller and probably is the reason why no significant differences were found for this case. This case however is taken into account because of the similar strong patterns in relation to case 1 and 2, although based on visual inspection. For analysing the data, we chose for a conservative approach because of the small sample size and possible violations of assumptions. This was done by analysing more variables in one analysis which reduced the degrees of freedom and thereby decreasing the chance of finding false positive results.

Table 3

Short Description of case 1, 2 and 3

Case 1	Case 2	Case 3
A train driver called in the impression he had hit a person and reported he was going to confirm the hit. The rail signaller informed co-workers and started to inform other train drivers to stop their train or slow velocity as prescribed. After a few minutes, the train driver reported he could not find anything and that it must have been something else. The rail signaller informed co-workers and train drivers that they could start driving again. The short time it took to stop and continue again, caused over a half an hour delay involving all trains on the trajectory.	The rail signaller was informed by mail that trains need to reduce velocity between a certain trajectory to a maximum of 40 km/h. Colleagues were informed and all approaching trains for this trajectory were called and informed according to procedures.	A train was standing still on a station with a difficult passenger. The police was on-site, but the train had to wait for the situation to be resolved. The train was standing nearby the level crossing of the station and caused the level crossing signals to malfunction (as it should). Because of the malfunctioning, all approaching trains were informed that they had to pass the level crossing with a designated speed below 10km/h.

IWS results

The average IWS during the day of the main event was "minimal effort" with a small deviation (M = 2.06; SD = 1.13). The IWS pattern for the cases further analysed have a stretch lasting 10 or more minutes. The IWS pattern for the two cases are presented in Figure 6. For further analyses, the IWS pattern will be used as a reference for the other methods and a distinction will be made between low IWS (0-2) and high IWS (3-9).

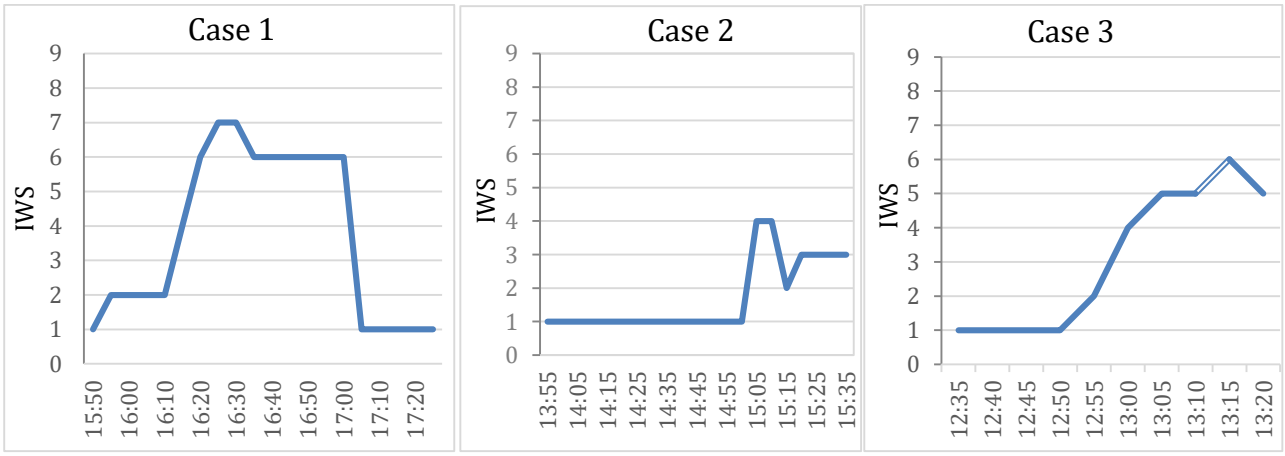


Figure 6. IWS scale 1 to 9 for cases one, two and three.

EDA results

The EDA data were visually inspected for any non-responders. All participants that seemed to provide usable EDA data were further analysed. Statistical analyses were conducted for the two cases using a MANOVA, comparing different EDA measurements (SCR, Amplitude, SCL) during the period of high IWS with the corresponding measurements during a period of low IWS. We found significant differences between periods of high and low IWS for all three measures in Case 1 (Figure 7). First of all, we found the SCR to be significantly different for periods of high IWS and for periods of low IWS ($F(1,18) = 8.58$, $p < .009$). The SCR signals occurred significantly more frequently for periods with a high IWS ($M = 87.4$; $SD = 18.96$) than for periods with a low IWS ($M = 63.60$; $SD = 17.68$). The amplitude was significantly ($F(1,18) = 8.59$, $p < .009$) higher for periods with a high IWS ($M = 27.73$ µS; $SD = 7.90$) than for periods with a low IWS ($M = 17.08$ µS; $SD = 8.33$). Also the SCL was significantly different for periods of high compared to periods of low IWS ($F(1,18) = 11.18$, $p < .004$). Again, the SCL was significantly higher for periods of high IWS ($M = 25.94$ µS; $SD = 5.37$) than for periods with low IWS ($M = 17.27$ µS; $SD = 6.19$). These results show that the three EDA measures can discriminate between high and low IWS in Case 1.

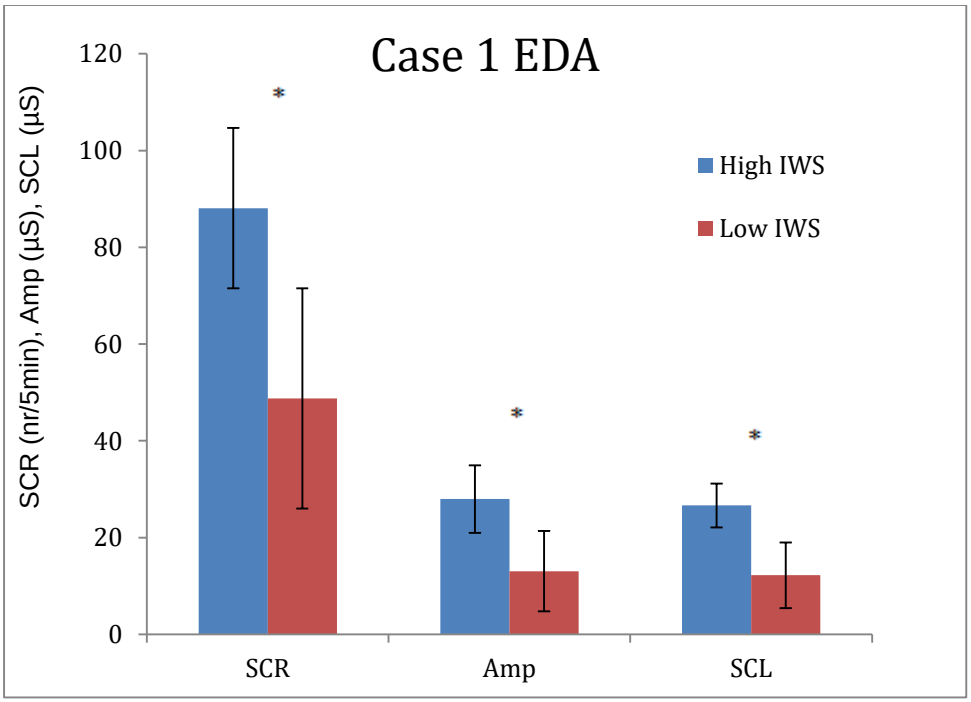


Figure 7. Average number of SCR, Amplitude (average µS) and SCL (average µS) for case one for high and low IWS. Significant differences are indicated with (*).

For case 2, only the SCL was significantly different (Figure 8; $F(1,19) = 1.66$, $p < .02$) for periods of high IWS ($M = 0.05$ µS; $SD = 0.08$) compared to periods of low IWS ($M = 0.65$ µS; $SD = 0.62$). The results for the SCR and Amplitude were not significant for case 2. Moreover, the effect for SCL in case 2 is incongruent with the results of case 1. In case 2, the SCL is significantly higher for periods with *low* subjective workload.

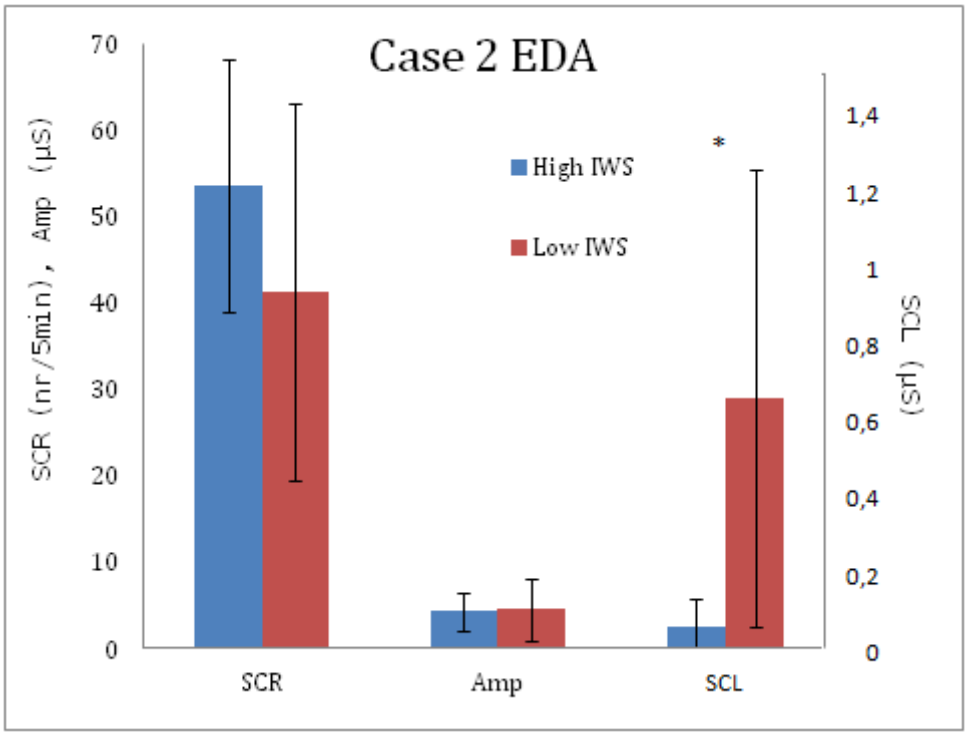


Figure 8. Average number of SCR, Amplitude (average µS) and SCL (average µS) for case two for high and low IWS. Significant differences are indicated with (*).

For case 3 (Figure 9), a similar pattern as in case 1 was found. The error bars between High and Low IWS do not overlap based on visual inspection and seem to differentiate strongly just as in case 1.

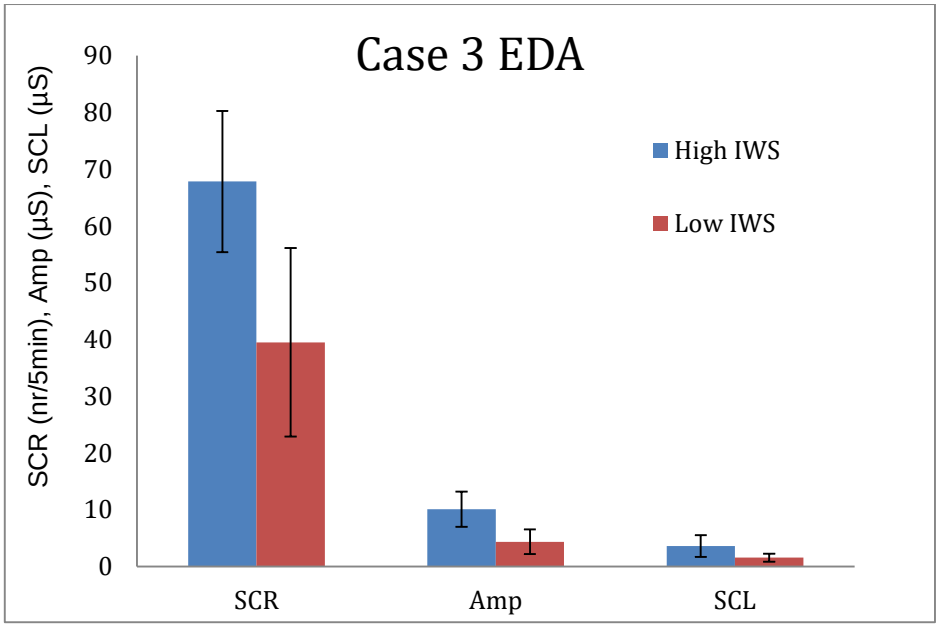


Figure 9. Average number of SCR, Amplitude (average µS) and SCL (average µS) for case three for high and low IWS.

BTL results

To corroborate the scoring system used, two of the researchers rated a sample of half an hour of observations. The inter-rater reliability was 85% (Cohen's kappa = 0.82) in number of seconds per behaviour.

For the behavioural observation results, we performed a similar MANOVA comparing the four BTL categories, number of switches between behaviours (Figure 10) and observed behaviours during periods of high IWS with the corresponding measurements during periods of low IWS.

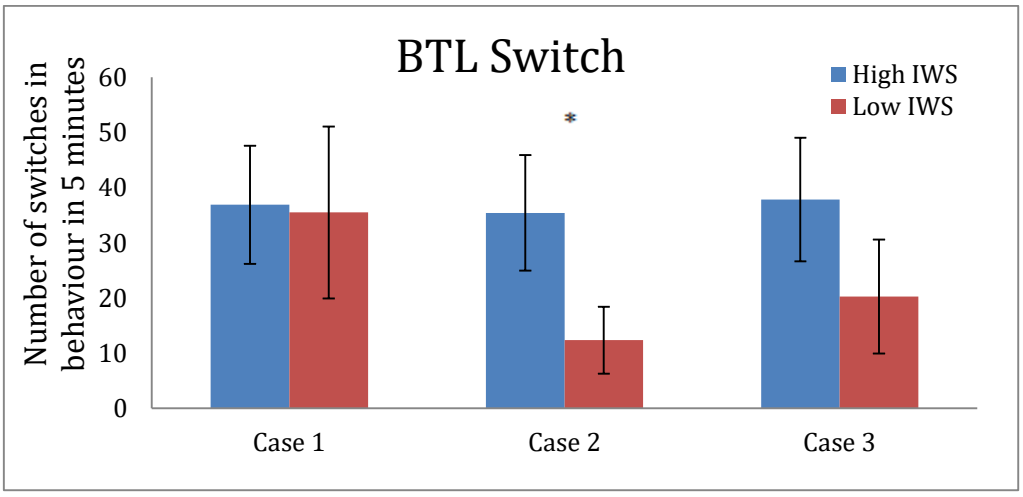


Figure 10. Average number of behavioural switches in 5 minutes intervals for case one, two and three for high and low IWS. Significant differences are indicated with (*).

For case 1 the factor communication differs significantly between periods of high and low IWS ($F(1,18) = 4.74$, $p < .04$). When looking at the subcategories of case 1 (Figure 11) for communication, we see that there is a significant difference for communication through telephone with a train driver ($F(1,18) = 10.70$, $p = .004$). This means that this behaviour occurs more during periods of high IWS ($M = 87.73$s (out of 300); $SD = 75.38$) than during periods of low IWS ($M = 8.86$s (out of 300); $SD = 11.48$). Also the local communication with the decentralized traffic manager was significantly different ($F(1,18) = 4.54$, $p = .05$) during periods of high IWS ($M = 9.53$s (out of 300); $SD = 13.52$) compared to periods of low IWS ($M = 0.38$s (out of 300); $SD = 1.20$). This means that communication through the telephone with a train driver and local communication with a decentralized traffic manager are significantly higher in a high IWS situation than in a low IWS situation.

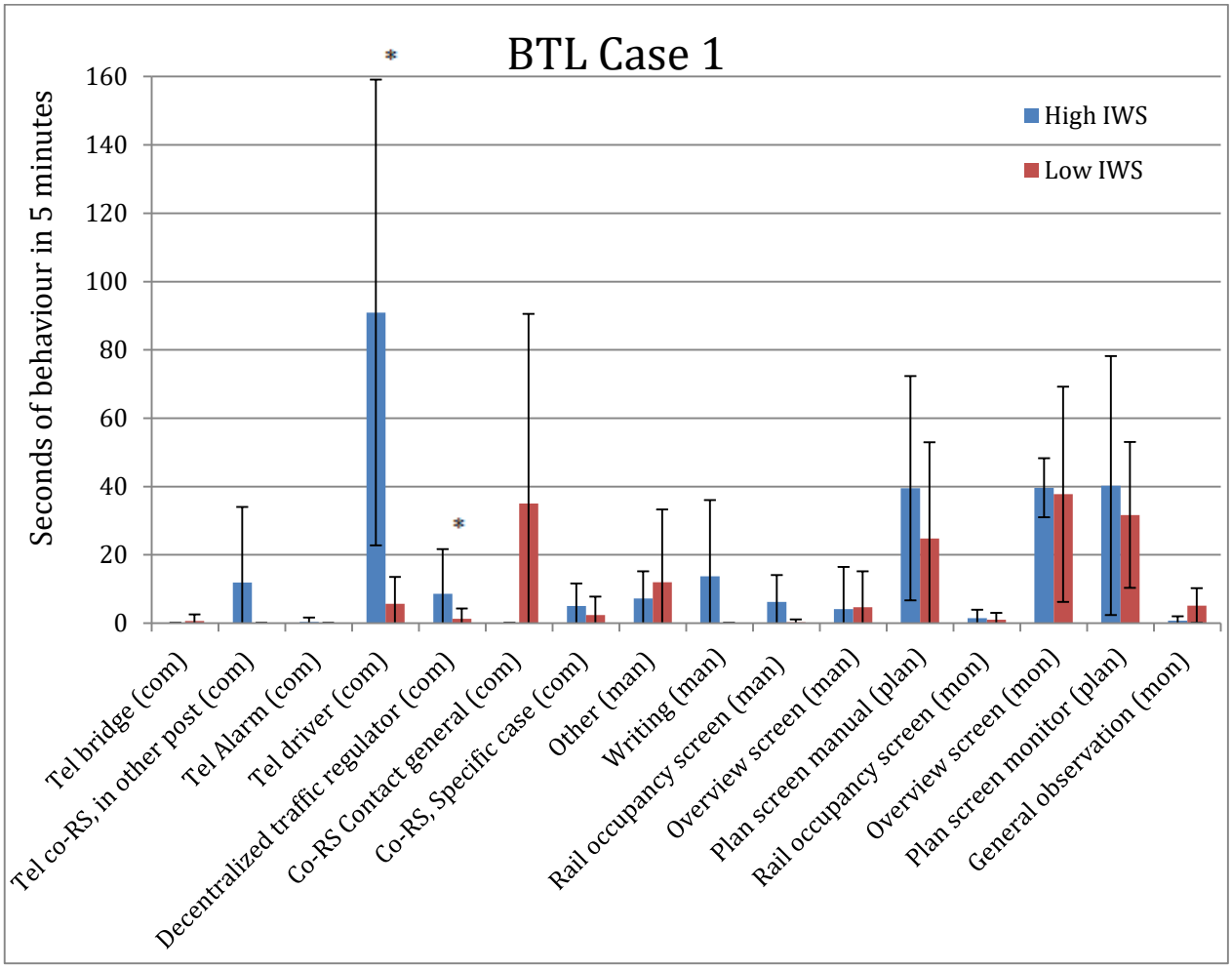


Figure 11. BTL observed behaviours for case one for high and low IWS. Significant differences are indicated with (*).

For case 2 the four BTL categories communication ($F(1,19) = 17.85$, $p < .001$), manual ($F(1,19) = 11.23$, $p < .003$), planning ($F(1,19) = 5.85$, $p < .05$) and monitoring ($F(1,19) = 21.70$, $p < .001$) were all significantly different between high and low IWS. Also, the number of switches between behaviours was significantly different for high and low IWS (Figure 10; $F(1,19) = 36.73$, $p < .001$), with more switches in 5 minutes for high IWS ($M = 35.43$; $SD = 11.33$) than for low IWS ($M = 12.36$; $SD = 6.30$). On the behavioural level (figure 12), communication through telephone with a train driver was significantly different ($F(1,19) = 10.36$, $p < .005$) for high IWS periods ($M = 63.7$s (out of 300) ; $SD = 76.10$) compared to low IWS periods ($M = 0.00$s (out of 300); $SD = 0.00$). Also, local (case specific) communication with colleagues was significantly different for high and low IWS, ($F(1,19) = 8.08$, $p < .01$), with more communication in high IWS periods ($M = 16.22$; $SD = 21.94$) than in low IWS periods ($M = 0.00$s (out of 300) ; $SD = 0.00$). Further, manually writing ($F(1,19) = 10.68$, $p < .004$) was significantly different for periods of high IWS ($M = 16.96$ s (out of 300) ; $SD = 19.95$) compared to periods of low IWS ($M = 0.00$s (out of 300) ; $SD = 0.00$). Monitoring planning screen was also significantly different for high and low IWS, ($F(1,19) = 10.95$, $p < .004$), with more monitoring during high IWS periods ($M = 17.82$s (out of 300) ; $SD = 7.33$) compared to low IWS periods ($M = 7.27$s (out of 300) ; $SD = 6.68$). Finally, monitoring the overview screen was significantly higher ($F(1,19) = 19.59$, $p < .001$) during high IWS periods ($M = 38.89$s (out of 300) ; $SD = 18.05$) compared to low IWS periods ($M = 8.73$s (out of 300); $SD = 6.68$).

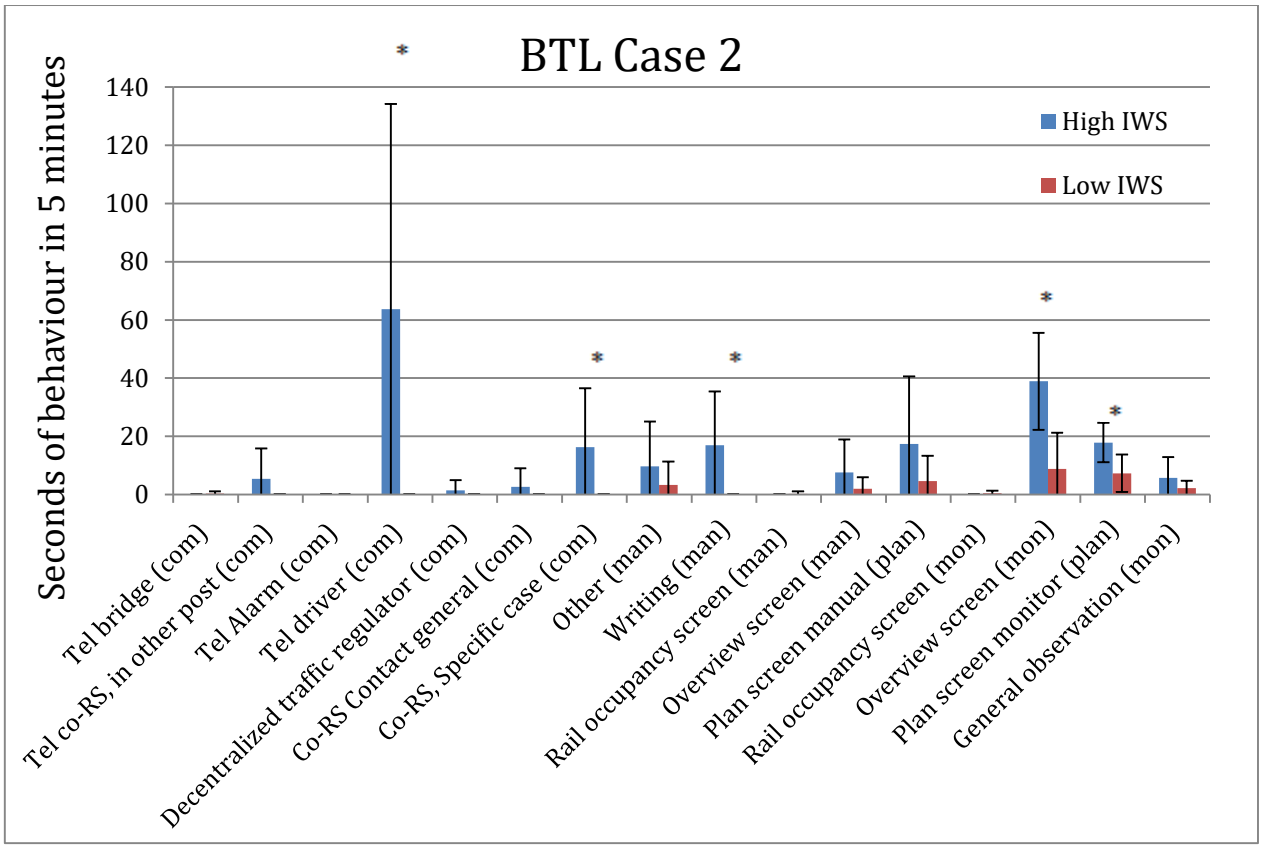


Figure 12. BTL observed behaviours for case two for high and low IWS. Significant differences are indicated with (*).

For case 3 again switch seems to differentiate for this case (Figure 10), for the specific behaviours (Figure 13), again communication with the train driver shows the largest absolute difference between high and low IWS in this dataset. Also writing and communication about a specific case seem to show differences based on visual inspection. In conclusion, the results from these three cases show that many specific behaviours were able to discriminate between periods of high and low IWS.

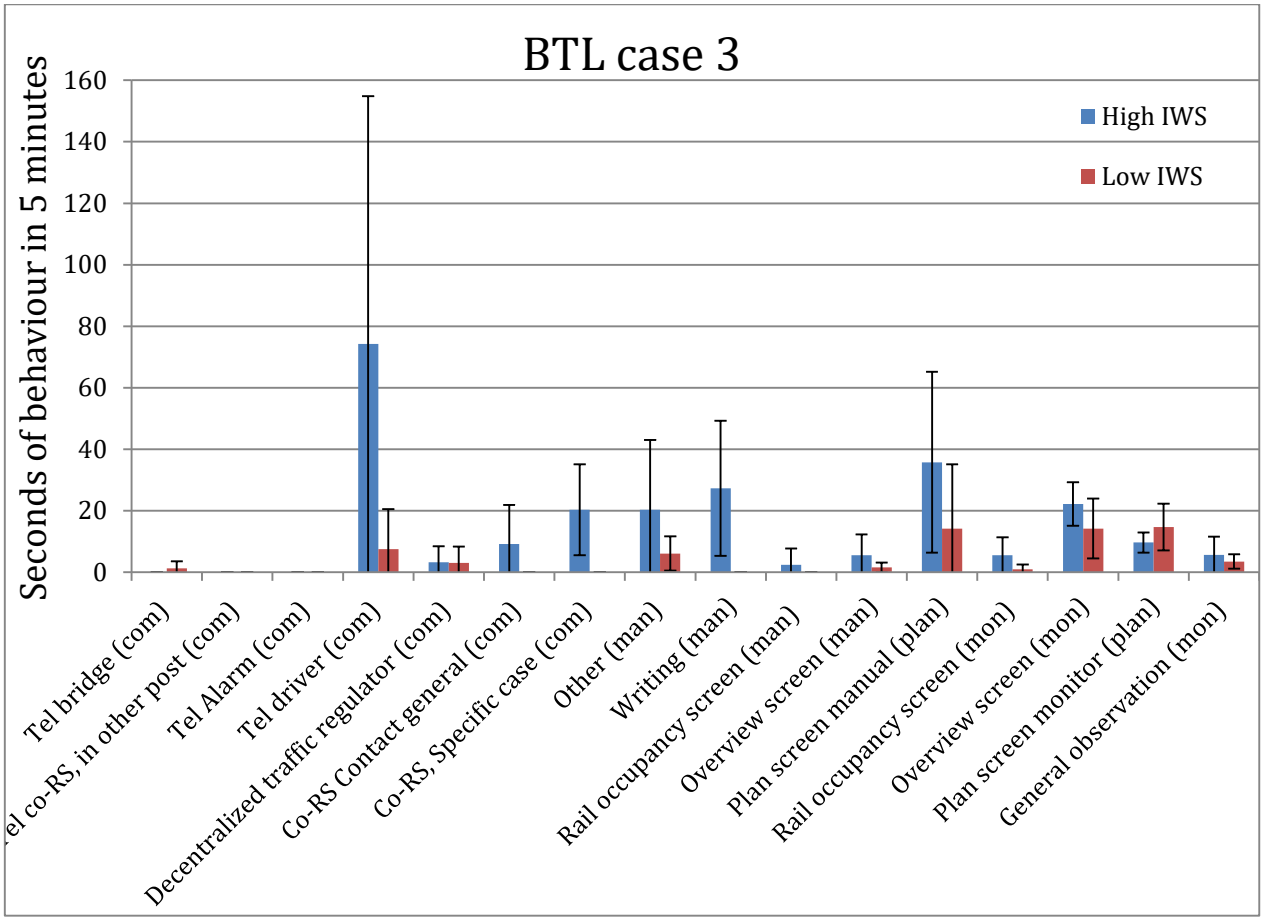


Figure 13. BTL observed behaviours for case three for high and low IWS.

XTL results

For the XTL data, we performed a MANOVA on low versus high IWS with five factors (Mon, Plan, Man, Com and Switch). For Case 1 (Figure 14, left panel), the factor communication differed significantly ($F(1,18) = 11.20$, $p < .004$) with more seconds per 5 minutes spoken trough telephone for

periods of high IWS (*M* = 0.25 XTL value/5 min; *SD* = 0.20) versus periods with low IWS (*M* = 0.03 XTL value/5 min ; *SD* = 0.06).

For case 2 (Figure 14, middle panel), the factor communication was significantly different ($F(1,19) = 8.58$, $p < .009$), with more seconds per 5 minutes spoken trough telephone for periods of high IWS (*M* = 0.25 XTL value/5 min; *SD* = 0.33), compared to periods of low IWS (*M* = 0.00 XTL value/5 min; *SD* = 0.00). Manual was also significantly different ($F(1,19) = 9.00$, $p < .006$), with more manual mutations in high IWS (*M* = 0.10 XTL value/5 min; *SD* = .13) compared to low IWS (*M* = 0.00 XTL value/5 min; *SD* = 0.00). Switch was also significantly different ($F(1,19) = 5.09$, $p < .04$) with more switches during periods of high IWS (*M* = 0.07 XTL value/5 min; *SD* = 0.06) than during periods of low IWS (*M* = 0.01 XTL value/5 min; *SD* = 0.04). In conclusion, these results show that communication, manual and switch cost discriminated between high and low IWS.

For case 3 (Figure 14, right panel) Switch and communication seem to support findings from case 1 and 2.

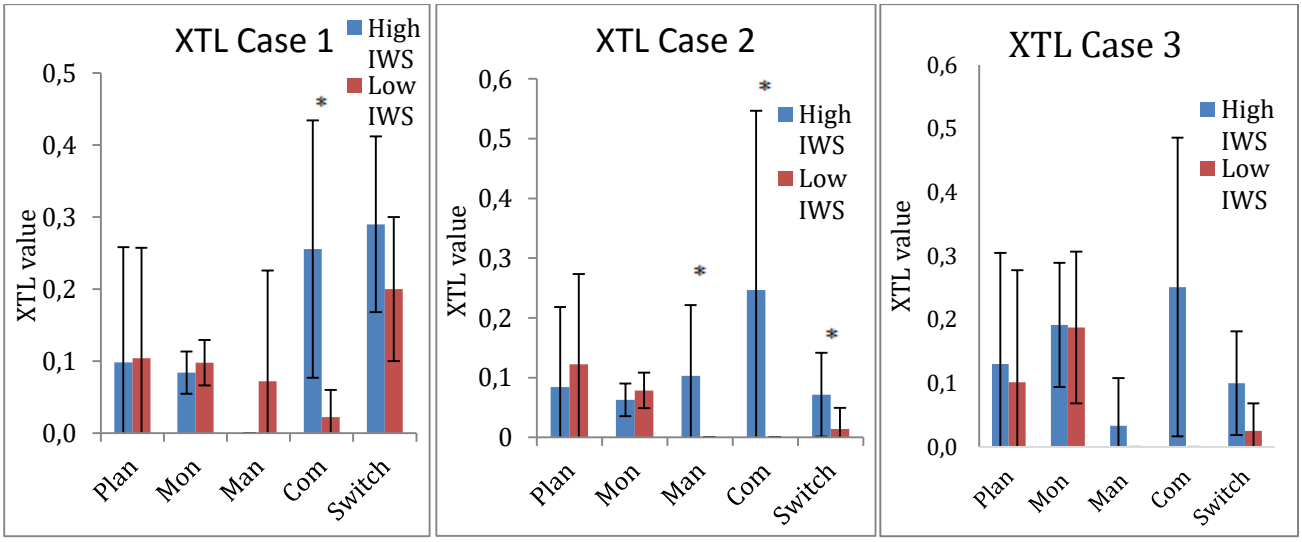


Figure 14. XTL parameters and switches for case one, two and three for high and low IWS. Significant differences are indicated with (*).

Discussion

This study investigated whether the four workload measurements (IWS, XTL EDA and BTL) support each other in the identification of changes in observed workload, and whether the XTL algorithm based on system data can be confirmed and complemented. First, the findings from the EDA, BTL and XTL measures will be discussed. Then the measures will be compared in a general discussion and limitations and implications are discussed.

Workload measurements

The results show that EDA is a good discriminator between high and low IWS values in case 1 and a trend for case 3, which is in line with the consensus that electrodermal activity is an online continuous physiological measure of workload that at least reflects a general measure of arousal or stress (Healey & Picard, 2005). This effect was not found in case 2 except for the SCL which was in the opposite direction than expected, but small. This discrepancy in case 2 could be explained by the smaller change of IWS, which did not pass the physiological arousal or stress threshold. Estes (2015) suggests that the subjective workload is better described as an s-curve, in which a low increase in task load leads to a low increase in experience workload in the beginning. Later, if the task load increases further, the experienced workload might also increase with large jumps, depending on available cognitive resources and strategies. This means that there is no gradually and linear increasing workload, but a low workload and then a relative sudden rise to high workload because there are no more extra cognitive resources or good alternative strategies available. This could mean that medium workload in IWS should be treated as low or high workload in the other measures. In this case, professional rail signallers who indicated medium experienced workload did not for example cause a large physiological response because they had extra cognitive resources available. This could explain that the IWS measure is not able to discriminate mediocre workload in relation to the EDA results. This complements a statement of Luczak and Göbel (2000) who state that a non-linear relationship would match be stress and strain demands. This is because of the amount of task load must first surpass a low minimum effort on the strain side to cause arousal. These findings however are contradicted by findings of Collet, Averty and Dittmar (2009) who found sort of a linear relation between task load and EDA measurement in professional air-traffic controllers. So further experiments are necessary to conclude this. Nonetheless EDA measurement results show that the EDA seems to be a promising method to differentiate between high and low subjective workload in rail signallers. The method is not intrusive with their work and it is theoretically possible to process the data in real time (although this was not the case in this research). If the findings are confirmed in other studies, EDA seems a good method to measure subjective workload in a non-intrusive way. However, for introduction of EDA measurements on the work floor or for wider applications, some ethical aspects of physiological data and feedback mechanism should be addressed (Fairclough, 2009). Privacy and autonomy of the rail signaller could be compromised if measures become obligatory or when it is not clear where the data are stored, who has access to the data and how or for what purposes the data are used.

BTL shows that different behaviours occur with high versus low IWS. Mainly the category

"communication" seems to be important. Looking at different behaviours, "telephone communication with train driver" and "contact with the decentralized rail traffic controller" came back in all three cases. When compared with the XTL, these effects reoccurred partially. The effects on the XTL, however, are less pronounced. The reason for this might be that, in the XTL, no distinction was made between with whom the telephone communication took place. This information seems to be important for interpreting these results, considering the BTL data. Also the factors "manual" and "switch cost" seem to differentiate between high and low workload. The BTL observations show that there is a high correlation between subjective and behavioural patterns, but that this highly depends on the behaviour in combination with other factors. For example, the same behaviours (communication/calling on the telephone) can have a different impact on experienced workload if the context or communicating partner is different. For the XTL, it would be desirable to make a differentiation for different categories or interactions in the socio-technical system. For example, calling with a train driver has probably more impact on experienced workload than calling with a bridge operator.

Finally, The XTL formula in the investigated cases shows a differentiating ability in both communication and manual actions. This shows that the XTL and in particular the parameters "communication" and "manual" could differentiate between high and low workload. However, for manual actions, the effects are not congruent and should be examined in the context of more cases. Also switch cost in XTL seems to show differentiating effect between high and low IWS (although not significant in case 1). For case 1 this could be explained by a lag of the IWS/XTL. If the last high IWS for case 1 is removed, the XLT switch cost is also significant. Also case 3 seems to support the differentiating effect for the factors communication and switch. The XTL could be improved in further research by differentiating the input data for XTL for the different categories. In this way, for example, a distinction could be made in who the other party in a telephone call is. These steps will make the XTL more sensitive and will create a better match between performance and experienced workload. The findings show that system information can be used to give some indication of experienced workload in rail signallers.

General discussion

This study investigated whether the four workload measurements (IWS, XTL, EDA and BTL) supported each other in the identification of changes in observed workload, and whether the on system data based XTL algorithm introduced by Siegel & Schraagen (2014) could be confirmed and complemented. Overall, the current research showed that real-time observation of subjective measures using IWS and XTL is feasible and can be corroborated by EDA and behavioural observation. The

results do show, however, that the measures show a better match when workload is very high then when the workload is more moderate. However, further research and specifications are necessary to determine and validate which of the system's data have a high predictive validity and which do not, considering the small number of cases. In terms of weak resilience signals the measures seem to be sensitive to larger increased workload, and are an objective addition to confirm the increased subjective workload.

This study showed that not only system information in car driving can be used to predict or measure workload in car drivers (Zeitlin, 1998; Ohm & Ludwig, 2013). But that it is also possible to use system information to predict and monitor workload in a professional rail setting. This study could therefore also contribute to measuring workload in other socio-technical systems where high trained professionals are operational and workload is an important factor, for example in air traffic control. One drawback of this study is that perhaps not enough attention was given to make a clear distinction between the stress and strain side (Schlegel, 1993). Perhaps in further research each executed behaviour or parameter could be described in terms of task load and an impact factor on the experienced workload in different situations and contexts. For example, calling with a train driver has the same task load as calling a co-rail signaller (picking up phone, dialling, talking) but is experienced as a higher task load. Probably, because of the context, the situation could be urgent and wrong information can have devastating consequences. So if other components in the socio-technical system, which mostly determine the situation, could be incorporated in the data collection the system data and XTL algorithm could be more sensitive and more discriminative between different situations and levels of workload. To even further refine the algorithm, machine learning could be applied as done before in the research of Tango, Minin, Tesauri & Montanari (2010) who refined a workload model for car drivers using system data and machine learning.

For the field of resilience engineering and specifically weak-resilience measuring, this study validates that system information can be used to detect differences in relative workload changes and could be used to monitor weak resilience signals in a continuous way without bothering or increasing workload on rail signalers in any way. This method presents an elegant way to gather relevant information which can be used to give rail signalers more insight about the workload of the shift and could enhance insight and guide feedback processes (Siegel & Schraagen, 2015). Enhancing the opportunity to collectively learn from situations by sharing knowledge in the rail control post and other partners in the socio-technical rail system, and thereby increasing resilience of the whole system (Hollnagel, 2009).

Acknowledgement

After conducting this study and writing this master thesis, I would like to thank the rail post Alkmaar for their openness and cooperation. Also I would like to thank TNO for providing assistance with the measuring equipment and the XT-observer. Matthijs Noordzij for his help with the Affectiva QTM sensor and for the EDA data processing. And last, but definitely not least, I would like to thank Jan Maarten Schraagen for the critical and thorough feedback and of course Willy Siegel for the support, feedback and guidance in this study, the paper and in completing this master thesis.

References

Amalberti, R. (2001). The paradoxes of almost totally safe transportation systems. *Safety Science, 37*(2-3), 109-126. doi: 10.1016/S0925-7535(00)00045-X

Bakeman, R. (2000). Behavioral observation and coding. *Handbook of research methods in social and personality psychology* (pp. 138-159). Cambridge University Press, New York.

Belmonte, F., Schon, W., Heurley, L., & Capel, R. (2011). Interdisciplinary safety analysis of complex socio-technological systems based on the functional resonance accident model: An application to railway traffic supervision. *Reliability Engineering & System Safety, 96*(2), 237-249. doi:10.1016/j.ress.2010.09.006

Benedek, M., & Kaernbach, C. (2010). Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology, 47*(4), 647–658. doi:10.1111/j.1469-8986.2009.00972.x

Boucsein, W. (2012). *Electrodermal activity* (2nd ed.). New York, USA: Springer. doi:10.1007/978-1-4614-1126-0

Brookhuis, K. A., & de Waard, D. (2011). Measuring physiology in simulators. In D. L. Fisher, M. Rizzo, J. K. Caird, & J. D. Lee (Eds.), *Handbook of Driving Simulation for Engineering, Medicine and Psychology* (pp. 17–1 – 17–10). CRC Press. Retrieved from http://worldcat.org/isbn/9781420061000

Collet, C., Averty, P., & Dittmar, A. (2009). Autonomic nervous system and subjective ratings of strain in air-traffic control. *Applied ergonomics, 40*(1), 23-32.

Cook, R., & Rasmussen, J. (2005). "Going solid": a model of system dynamics and consequences for patient safety. *Quality & Safety in Health Care, 14*(2), 130-134. doi:10.1136/qshc.2003.009530

Estes, S. (2015). The Workload Curve Subjective Mental Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(7), 1174-1187.

Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interacting with computers*, *21*(1), 133-145.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183. Retrieved from http://humanfactors.arc.nasa.gov/groups/TLX/downloads/NASA-TLXChapter.pdf

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

Hollnagel, E. (2009). The four cornerstones of resilience engineering. In C. P. Nemeth, E. Hollnagel, & S. Dekker (Eds.), *Resilience Engineering Perspectives. Volume 2: Preparation and restoration* (pp. 117–134). Farnham, Surrey: Ashgate Publishing Limited.

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

Luczak, H., & Göbel, M. (2000). Signal processing and analysis in application. *Engineering psychophysiology: Issues and applications*, 79-110.

Madni, A. M., & Jackson, S. (2009). Towards a conceptual framework for resilience engineering. *IEEE Systems Journal*, *3*(2), 181–191. doi:10.1109/JSYST.2009.2017397

Neerincx, M. A. (2003). Cognitive task load analysis: allocating tasks and designing support. In E. Hollnagel (Ed.), *Handbook of cognitive task design* (pp. 283–305). Mahwah, NJ: Lawrence Erlbaum Associates.

Obinata, G., Tokuda, S., Fukuda, K., & Hamada, H. (2009). Quantitative evaluation of mental workload by using model of involuntary eye movement. In *Engineering Psychology and Cognitive Ergonomics* (pp. 223-232). Springer Berlin Heidelberg.

Ohm, C., & Ludwig, B. (2013). Estimating the Driver's Workload. In *KI 2013: Advances in Artificial Intelligence* (pp. 130-139). Springer Berlin Heidelberg.

Pickup, L., Wilson, J. R., Norris, B. J., Mitchell, L., & Morrisroe, G. (2005). The Integrated Workload Scale (IWS): a new self-report tool to assess railway signaller workload. *Applied Ergonomics*, *36*(6), 681–693. doi:10.1016/j.apergo.2005.05.004

Poh, M.-Z., Swenson, N. C., & Picard, R. W. (2010). A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE Transactions on Bio-Medical Engineering*, *57*(5), 1243–1252. doi:10.1109/TBME.2009.2038487

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling

procedure for measuring mental workload. *Advances in psychology*, *52*, 185-218.

Rubio, S., Diaz, E., Martin, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology-an International Review-Psychologie Appliquee-Revue Internationale*, *53*(1), 61-86. doi:10.1111/j.1464-0597.2004.00161.x

Schlegel, R. E. (1993). "Driver Mental Workload." In Automotive Ergonomics, edited by B. Peacock, and W. Karwowski, 359–382. London: Taylor & Francis.

Siegel, A. W., & Schraagen, J. M. C. (2014). Measuring workload weak resilience signals at a rail control post. *IIE Transactions on Occupational Ergonomics and Human Factors*, *2*(3-4), 179–193. doi:10.1080/21577323.2014.958632

Siegel, A. W., & Schraagen, J. M. C. (2015, June). Can team reflection of rail operators make resilience-related knowledge explicit? - an observational study design. In *6th Resilience Engineering Symposium, Lisbon,* (pp. in production).

Tango, F., Minin, L., Tesauri, F., & Montanari, R. (2010). Field tests and machine learning approaches for refining algorithms and correlations of driver's model parameters. *Applied ergonomics*, *41*(2), 211-224.

Tsang, P. S., & Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*,*39*(3), 358-381. doi:10.1080/00140139608964470

Woods, D. D., Wreathall, J., & Anders, S. (2006, November). Stress-strain plots as a model of an organization's resilience. In *Proceedings of the 2nd Resilience Engineering Symposium, (*pp. 342-349).

Young, M. S., Brookhuis, K. a, Wickens, C. D., & Hancock, P.A. (2015). State of science: mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17. doi:10.1080/00140139.2014.956151

Zeitlin, L. (1998). Micromodel for objective estimation of driver mental workload from task data. *Transportation Research Record: Journal of the Transportation Research Board*, (1631), 28-34.