Final Project

Measuring Workload at a ProRail Control Post

Jaldert van der Werf
s.j.vanderwerf@student.utwente.nl

Faculty of Electrical Engineering, Mathematics and Computer Science
University of Twente

August 22, 2014

Contents

1 Introduction

The resilience of an organization is the ability of the organization to cope with disruptions or adapt to events that are of influence on its operations. An organization is resilient if it is able to handle these disruptions or events in such a way that it can continue its operations. Changes in the resilience of an organization can be identified as weak resilience signals (WRS). Siegel and Schraagen[1] propose a model to identify WRS at the Dutch railway system, which is an example of socio-technical system. Three boundaries are taken into account to identify a WRS; the performance, safety and workload boundaries. The boundaries are aspects where pressure occurs that influences the resilience of the organization. The performance boundary is taken into account for the railway sector since it is judged on its performance such as train delays. Safety is a boundary since it is of paramount importance that all operations are handled safely. Workload is a boundary since it encapsulates the influence of the human factor within an organization.

Siegel and Schraagen provide a theoretical model for identifying WRS, the focus for this research is on the workload boundary. To identify a WRS at the workload boundary it is necessary to measure workload changes real-time in a control room setting of a complex socio-technical system. Real-time measuring of workload in an operational environment of such a system is the main goal for this final project. The focus is on measuring workload changes with regarding to WRS since this has not be done in this context and it provides the first step in Siegel and Schraagen's model for identifying WRS.

The focus is a control room setting, which is an example of a socio-technical system since technical systems and human operators are the key concepts. Changes in the workload of an operator are a potential indicator for a change of the resilience of the system. These workload changes are an indicator since changes in workload result in pressure on the workload boundary. Whenever the workload of an operator changes it is the result of a change of the state of the system that is acted on by the operator.

Workload measurements for this research are performed at ProRail control post Zwolle. ProRail is the company responsible for rail logistics in the Netherlands and has several control posts over the country. Train traffic is managed and controlled by rail signalers (train traffic controllers) at these control posts. Workload of the rail signalers of post Zwolle is measured real-time in their operational environment.

1.1 Research questions

Measuring workload changes in a real-time setting is the main goal. First workload is defined since you need to know what is meant by workload and how to measure it. A methodology to measure workload in a real-time setting is created based on the workload definition for rail signalers. The methodology forms the basis for the measurement sessions in a real-time setting at a control room. This leads to the following research questions.

The main research question is:

- *How can one measure workload of a socio-technical system in a real-time setting?*

To answer this question the following sub questions need to be answered:

- *Are there relevant existing methods to measure workload in socio-technical (control room) setting?*
- *Can these methods be applied to measure workload at a ProRail control post or how can these methods be adjusted or extended to be used in this setting?*
- *How can one detect changes in workload by measuring workload in a real-time setting?*

1.2 Workload definition

Workload measurements take place in a control room setting which is operational for 24 hours a day. The typical workload in such a setting is quite low since most systems are automated and operators are monitoring for the majority of time. This results in a stable workload since the main activity remains the same throughout the day. Changes in workload are therefore of interest. A change in workload is most likely caused by breaking the monitoring routine which requires effort from the operator. Increases in effort caused by changes in workload are of interest since it is an indication that an activity or event of interest is happening. By this approach workload is defined by the complexity of tasks and the amount of work performed by operators. These indicators for workload are about the cognitive capacities of the operator and can therefore be defined as cognitive workload. When operator tasks are more complex or the amount of work increases it requires more from the operator's cognitives capacities. The complexity of tasks and amount of work form two indicators for measuring the cognitive workload.

Determining workload based on cognitive capacities is one way of defining the workload of an operator. His mental state or perception of work are another type of workload, since these indicators are not directly linked to the operators cognitive capacities. The subjective state of an operator determines the subjective workload. To measure this subjective workload you need to record the perception of the operator. Besides the cognitive and subjective there is the physical state of an operator that translates to a third type of workload. The physical state of an operator can be used to determine the physiological workload. Measuring this type of workload can be done by measuring the physical state of an operator during their job.

The cognitive, subjective and physiological workload types are discussed in related work (section 3). Existing methods for measuring these types of workload are discussed. Measuring these three types of workload provides a versatile view on the workload of the rail signalers at the control post. The methods to measure these types of workload in similar control room settings are discussed, there is determined how the discussed methods can be applied to the situation at the ProRail control room. Combining existing methods is the first step, since a combination of these methods alone is not sufficient to measure workload at a ProRail control post. These methods are used as a basis for measurement sessions at ProRail and are adjusted or extended before they are applied in this real-time setting.

1.3 Requirements

The measurement sessions are performed in an operational environment and are integrated in a real-time setting oppose to a test setting. Performing the measurements in an operational environment has several limitations. The most important one is that there should be minimal interference with the routine of rail signalers, since they are doing their daily job. No interference with rail signalers normal routines is the main requirement during the measurement of workload, there should be no difference in the way they execute their jobs. Some input from rail signalers is required though, for instance for measuring the subjective workload. The next requirement is therefore to minimize the amount of input from rail signalers. Measurements are performed using software tooling and other peripherals that need to be integrated in the operational environment. A requirement is that the set up time is limited, preferably within five minutes. Once the measurements are active there should be no need for the researchers to adjust settings or to interfere with the measurements in another way, all systems should be able to run for longer periods without a need for maintenance.

This results in the following requirements to integrate the workload measurements in a real-time setting:

- [**R1**] No interference with normal routine of rail signalers.
- [**R2**] Minimize required input from rail signalers (no more than one action per five minutes).
- [**R3**] Limit the set up time as much as possible (within five minutes).
- [**R4**] Keep systems running, no input from the researches should be required once the system has been set up.
- [**R5**] Keep systems running for longer periods, ideally 24/7.

1.4 Overview

The workload measurements take place at ProRail, the setting at ProRail in which rail signalers execute their job is described first. By defining the setting at the control post it is possible to determine what methods for measuring workload changes are applicable. There is looked into existing methods for measuring workload at control room settings which are applicable to the setting at ProRail as well. This consists mainly of studies for workload measurements in control room settings. Numerous studies in this field exist, the methods used in these studies are of interest for this research as well when they are applied in a similar setting. Existing methods are discussed in the related work section, the requirements are used to determine if a method can be used at ProRail. A method needs to be adapted or adjusted if the requirements are not met. In the end it results in applicable methods for measuring workload changes in a real-time setting at ProRail. Most important is to adjust a method in such a way that it is usable in this operational environment.

A methodology for measuring workload at the control post using the adapted methods is created (section 6). Workload measurements are conducted based on this methodology. Results of the measurement sessions are discussed in section 8, these include the measured changes of the different types of workload. The integration of the workload measurements in the real-time setting at ProRail and other research questions are handled afterwards (section 10).

2 Setting

The research was conducted at ProRail control post Zwolle. ProRail is responsible for rail logistics in the Netherlands and has thirteen control posts over the country. Each control post handles train traffic in a specific region, post Zwolle handles train traffic in the north-eastern part of the Netherlands. A ProRail control post is a control room setting and an example of a socio-technical system due a high level of interaction between humans and the system. This setting was therefore suitable for measuring workload changes with regarding to weak resilience signals.

2.1 ProRail post Zwolle

Control post Zwolle is operational for 24 hours a day, 7 days a week. There are between 10 and 20 professionals active at each moment depending on the time and circumstances of the day. Rail signalers (train traffic controllers) were the group of interest for this research. Each rail signaler has a work station from which he manages train traffic for a number of railway stations and intermediate tracks. Rail signalers are responsible for all train traffic on the rail infrastructure, their tasks of managing train traffic consist mainly of logistics and safety tasks.

Post Zwolle was planning a reorganization to improve their performance and efficiency, tasks of rail signalers were divided differently because of this reorganization. In the original situation rail signalers were responsible for both the logistics and safety tasks. A rail signaler $\mathrm{RS_i}$ managed train traffic on a work station $\mathrm{WS_i}$, this included logistics tasks ($\mathrm{L_i}$) and safety tasks ($\mathrm{S_i}$) for this work station($\mathrm{WS_i}$).

Rail signalers were divided in two groups, a target ($\mathrm{G_T}$) and a reference ($\mathrm{G_R}$) group. The target group worked with split tasks after the reorganization and the reference group worked according to the original method (both safety and logistic tasks) during the complete research period. Tasks for rail signalers were split after the reorganization, meaning that a rail signaler handled safety or logistic tasks only.

The groups were divided based on corridors. A corridor is a larger area of rail infrastructure that consists of a number of adjacent areas from different work stations, see figure 1. The reasoning behind the reorganization was that collaboration in a corridor leads to more efficiency and therefore less calamities. To be implemented aspects of the reorganization at the time of research were an increase in rotation of rail signalers between work stations. This should enforce standardization in the way rail signalers operate a specific work station. Training will also be included to the routine in the future to enforce rail signalers to handle certain situations more efficiently.

Figure 1: Corridors for G_T and G_R

2.2 Splitting tasks

Splitting the logistics and safety tasks was the most structural change and had a potential influence on the workload. The target group G_T consisted of four rail signalers $RS_{1..4}$, three of them used to operate work stations ($WS_{1..3}$) with concerning logistic and safety tasks $L_{1..3}$ and $S_{1..3}$. One rail signaler was extra and available for assistance during calamities or rush hours. These four rail signalers handled the same work stations after the reorganization, the tasks were now split however. Safety tasks form the majority of work, so $RS_{1..3}$ handled safety tasks $S_{1..3}$ on $WS_{1..3}$. RS_4 handled logistic tasks $L_{1..3}$ on an extra work station (WS_4). Changes in the tasks for the G_T are depicted in figure 2.

Figure 2: Task deviation for G_T. Changes in deviation of $L_{1..n}$(logistics) and $S_{1..n}$ (safety) tasks between work stations $WS_{1..n}$ before and after the reorganization.

The research was therefore situated around the reorganization. The research consisted of two sessions of a week (Monday till Friday) the first being before the reorganization and the second after the reorganization. Both G_T and G_R worked with the combined safety and logistics tasks before the reorganization, after the reorganization G_T worked with split tasks whereas G_R continued to work with combined tasks.

The target group operated the northern corridor of post Zwolle and had four rail signalers, the reference group consisted of three work stations. Rail signalers in this group performed both logistics and safety tasks before and after the reorganization. A separate calamity work station was also available and could be used to assist during calamities. Both work stations from the target and reference group could be assisted from this extra work station during calamities.

3 Related work

To select applicable methods for measuring workload of rail signalers there was first looked into existing methods. Numerous methods for measuring workload at control room settings exist. There was a need to measure different types of workload as defined in the first section. These types of workload entail the effort and amount of work of rail signalers, as well as the mental and physical state of rail signalers. It was therefore necessary to define methods to measure these different types of workload. The first step was to look into existing methods that measure these types of workload in a control room setting, similar to the setting at ProRail. A combination of existing methods had to cover the different types of workload and the methods had to match or could easily be adapted to met the requirements from section 1.3. The resulting methods were used to measure workload changes of rail signalers.

Relevant methods are described and their applicability to the measurement sessions at ProRail is discussed in the next section. This provides an insight in the usability of these methods for this research and brings about the shortcomings of these methods with regarding to criteria applicable to ProRail.

3.1 Cognitive task load

Neerincx[2] described a model to measure cognitive task load. In this model he used task characteristics to model the performance and mental effort of the operators performing a specific task. The first characteristic is described as the *percentage time occupied*, which is the time that people are working on tasks that are required for their work. An important notion is that people should not be occupied more than 70 to 80 percent of the total time they are available. The percentage of time that people are occupied combined with the duration of tasks forms a classic way to assess workload.

A second characteristic is the *level of information processing* or the task difficulty. This is based on the idea that low level tasks are handled more or less automatically and therefore take less effort. Rasmussen[3] describes this concept in his Skill-Rule-Knowledge framework. Information is processed automatically with hardly any cognitive demand on a skill based level. The next level is rule based that consists of tasks that are executed using a fixed routine, typically consisting of re-occurring actions. These include tasks that can be performed when certain requirements are met, operators simply execute these tasks based on rules. Executing such tasks requires more cognitive demand compared to the skill based tasks. The third and final level is knowledge based, here the operator needs to analyze input for the tasks and come up with a solution. The knowledge base level applies particularly when an operator has to deal with a new situation.

The final characteristic that contributes to the load factor is *task-set switching*. This characteristic is especially applicable to complex tasks since these usually exists of a number of different sub tasks. Solving the greater task therefore requires different skills and knowledge simultaneously.

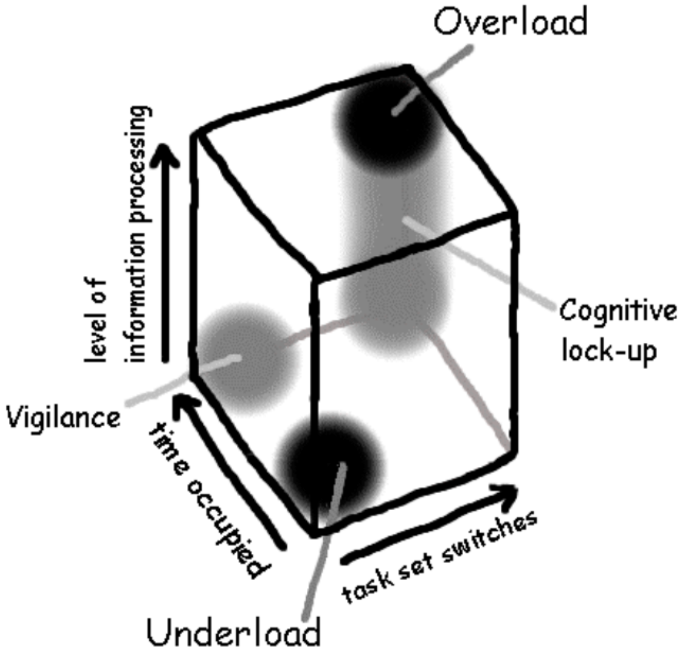

Figure 3: Cognitive task load model by Neerincx[2]

3.1.1 Creating a three dimensional load factor

The combination of task duration, level of information processing (difficulty) and switches form the cognitive task load. Neerincx used these three characteristics to create a three dimensional load space, see figure 3, in which human activities can be projected and have an indication of the cognitive load. The model defines states of task load that can occur. Overload occurs when the values for all three characteristics are high, it can occur at any give time. Underload on the contrary occurs when the values for the characteristics are low, temporary low values are not problematic and will not result in underload directly, underload occurs when the values are low for a significant amount of time. Another state of the model is vigilance in which problems increase in time. Vigilance causes a decrease in performance and stress due to specific task demands, such as the need to continuously paying attention on the task. Boredom can also be a sign of vigilance especially with high repetitive and homogeneous tasks. The final state of the task load model is cognitive lock-up which means that humans are incapable of adequately managing their tasks. Especially prioritizing between tasks is problematic since humans tend to stick to their current task even if a new task has a higher priority.

3.2 Stress-Strain model

Another approach is used in the stress-strain model by Woods[4]. This model is based on an analogy to the world of material sciences to measure the resilience of a socio-technical organization. Siegel and Schraagen[1] propose a resilience state model for a railway organization that is based on Woods' stress-strain model. They model resilience on the safety, performance and workload boundaries. The stress-strain model is used to measure the resilience of an organization by mapping applicable components of the model to measure workload for the railway sector.

Woods' model is plotted as a stress-strain plot known from material sciences. Stress is plotted on the y-axis and is depicted as demand (D), strain (S) is plotted on the x-axis and corresponds to the load analogous to the stretch of a material. Material sciences defines two regions of behavior, first their is elastic behavior. In the elastic region the demand (stress) and load (strain) increase uniformly, when the demand goes up the load rises in a similar fashion. Secondly there is a plastic region in which a material stretches non-uniformly until a failure point is reached.

The elastic region, referred to as the uniform response region, describes an organization with developed plans, procedures, training, personnel and related operational resources. The demand and load are balanced in this region, meaning that everything goes according to plan. The second region is characterized by non-uniform behavior. In material sciences this results in a non-uniform stretch or deformation of the material. In an organization it results in gaps in the ability to maintain a safe and effective production. These gaps are caused by a increasing demand and a load that cannot keep up with this rising demand. Plans and procedures need to be adapted to cope with this situation and to make sure it will not end up in failure. By adjusting strategies and acquiring additional resources it is possible to create a so called extra region. The extra region extends the capacity of an organization to cope with a increasing demand. This process can be repeated several times to avoid ending up in new failure point when these extra resources are exhausted. Figure 4 shows the stress-strain plot described in this section.

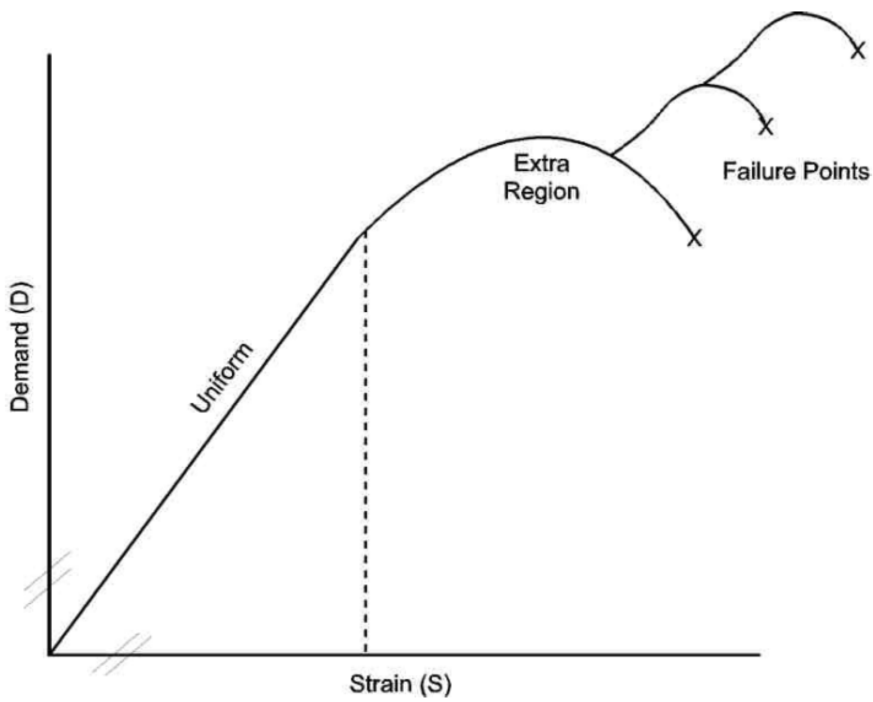

Figure 4: Demand (Stress) - Strain (Load) plot by Woods[4]

3.3 Integrated Workload Scale

Pickup developed a conceptual framework for mental workload[5]. Mental workload is described in this framework and is divided in three kinds of load; imposed, internal and perceived load. Imposed load is based on the task characteristics, internal and perceived load are based on individual characteristics. The framework is used by Pickup to create a new way to measure mental workload by developing an empirical assessment tool called *The Integrated Workload Scale (IWS)*[6].

The IWS scale is used by operators to rate their own workload. Self reporting by operators provides an accurate reflection of their workload, since operators are well aware of the situation they are coping with, the possibilities and limits of their equipment and they can estimate the workload for a situation based on previous experiences. One of the most widely used scales to measure workload is the NASA-TLX scale[7], but this scale is not usable in real-time in the field and not representative for rail signalers. Pickups IWS scale is specifically developed for rail signalers.

An multidimensional or unidimensional scale can be used to self report mental workload. For a multidimensional scale each aspect of workload needs to be rated separately whereas for an unidimensional scale the aspects are combined on one scale. The IWS scale is developed as an unidimensional scale since detecting peaks and troughs in the workload is the most important aspect for the rail sector. This is because the overall subjective state of the rail signaler is of interest and not subjective ratings of aspects of their work. Fluctuations in the global workload can be detected using this unidimensional scale plus it is also the most convenient way for rail signalers since they need to rate their workload on one scale only.

The IWS scale is based on the Instantaneous Self-Assessment (ISA) scale[8] since this scale was practical and very simple to use. Descriptors of the ISA skill were changed to terms that reflected the workload level of rail signalers more accurately. Changing the descriptors to terminology used by rail signalers should enhance the validity of the IWS tool. The new descriptors were gathered from interviews with rail signalers regarding workload. This approach let to typical phrases to describe workload such as ”struggling to keep up”. Two human factor researchers gathered a set of 47 phrases and terms used to describe workload. Rail signalers were asked to score these phrases and this let to a nine-point scale. This IWS scale is depicted in figure 5.

Figure 5: IWS scale for rail signalers by Pickup[6]

Mental workload of rail signalers is gathered by letting rail signalers rate their workload on the IWS scale. It is possible to see fluctuations in workload when rail signalers rate their workload at a fixed interval, see figure 6 for an example output with an interval of 2 minutes. The IWS scale was tested in a simulator environment where rail signalers had to provide an IWS score at an interval of 2 minutes and it was also tested in a real world situation where rail signalers provided a score at 5 minute intervals. IWS ratings provided an interpretation of the workload as experienced by rail signalers.

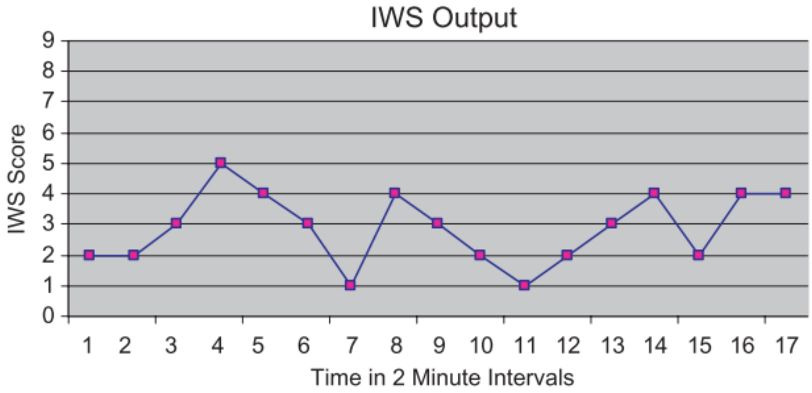

Figure 6: Typical IWS output over a 35 minute period by Pickup[6]

3.3.1 IWS in practice

The IWS scale has been used in practice by Wilms and Zeilstra[9] in a real life setting. They conducted a study in the Netherlands for ProRail and used the IWS scale to measure subjective workload at 26 workstations. The experiment was performed in one-hour sessions in which a researcher asked for IWS scores of rail signalers at 5 minute intervals. Furthermore researchers watched for seven predefined activities such as monitoring, communication activities, computer interaction and reading/writing activities. Besides these predefined activities they also registered environmental (disruptions, type of workstation, time period) and personal (age, experience, gender) variables. The activities, environmental and personal variables were used to calculate an objective workload which was compared to the subjective workload derived from the IWS scores. No correlation was found between objective and subjective workload. Wilms and Zeilstra do not have a good explanation for this, they point out that further research is required and that one should be cautious when comparing and interpreting results from different tools to measure workload. A changed approach such as longer measuring sessions of IWS and a different way to extract the objective workload provide opportunities. Wilms and Zeilstra had researchers in the field who where asking rail signalers to rate their workload. Interrupting rail signalers in their normal routine can be of influence on their subjective workload. One of the requirements for this research was therefore that rail signalers should not be distracted from their work.

3.4 Physiological methods

Gao[10] studied a number of different methods to measure physiological aspects to determine workload. The most significant physiological aspects are blink rate, blink duration, pupil size and heart rate variability (HRV) and were tested for sensitivity, validity and intrusiveness. The experiment consisted of 18 participants who had to perform a computerized emergency operation procedure (EOP) of a nuclear power plant. Workload measurements using the listed methods were gathered during the experiment. The main conclusions of this study were that blink rate is sensitive to both peak and overall heart rate. Pupil size is most sensitive to arousal caused by errors, blink duration increases during the task period regardless of the complexity of the EOP and heart rate variability measurements were able to distinguish tasks with different overall complexity. Gao used a model to integrate the six physiological methods to assess the overall workload accurately and reliably.

3.4.1 Heart rate variability

Heart rate variability (HRV) is a widely used method to measure workload. HRV is the variation in the duration of the R-R interval[1] which has become a popular measurement tool in various studies. Billman[11] provided a historic perspective of HRV usage. This historic overview starts around 300 BC when the Greek physician and scientist Herophilus first described heart rate based on pulses, it continues via Galen who wrote more than 18 books about heart rate around 170 AD and used the pulse for diagnosing diseases. The real break through was in the early 18th century when the invention of the Physician's Pulse Watch by John Floyer made is possible to assess heart rate more accurately. During this time it was also discovered that the pulse varied with respiration. Heart rate variability was recorded during the 19th century and took a giant leap with the invention of electrocardiography (ECG) in 1895 and digital signal processing techniques in the 1960's.

3.4.2 Real time detection workload changes

HRV was researched extensively especially due to it medical nature, but there are also numerous studies of HRV in relation to workload. Hoover[12] described a method to measure HRV to detect real time changes in workload. He created an experiment were participants had to play a video game (style of *Space Invaders*) and had to perform two tasks, an active 'shooting task' and a passive 'surveillance' task. Participants had to control a space ship. During the shooting task they had to fire at enemy ships, collect enough health and ammunition and needed to make sure they were not hit by enemy ships. During the surveillance task participants and enemy ships could not fire and it was not necessary to collect health or ammunition, it was only necessary to press the space bar when an specific enemy ship appeared. This set up created a different workload experience for both tasks, the tasks were also switched at random intervals during the experiment. HRV is inversely proportional to stress, a decrease in HRV is therefore an indicator of stress. The expected results were therefore that the active task would lead to a decrease in HRV and the passive task to an increase in HRV. Changes in HRV were very subtle, but the method was able to detect changes in HRV when the tasks were switched. Hoover stated that his approach has potential, the results should be handled carefully however. Detected changes in HRV were very subtle and could have been influenced by other factors as well.

3.4.3 Measuring HRV

Measuring the heart rate is the first step in determining the HRV. Once the heart rate, or R-R intervals are collected it is possible to determine the HRV in a number of different ways. HRV can be determined on the time or frequency domain[13][14]. Togo[15] described suitable methods for both the time and frequency domain. The SDNN (standard deviation over all normal-to-normal intervals) is a suitable method on the time domain. HRV is determined via the SDNN method by calculating the standard deviation of R-R intervals over a fixed periods of time, 5 minutes

[1] R is the upward deflection in an ECG, the interval between two consecutive R values is the R-R interval or heart period.

is often suggested. In the frequency domain the low-high frequency ratio (LF/HF) method is suitable. Calculating the LF/HF ratio can be done through a discrete Fourier transformation (DFT) by taking the first 2^n samples (this comes to the first 256 samples in a five minute interval). The LF is the spectral integral of frequencies between 0.04 Hz and 0.15 Hz whereas the HF uses frequencies between 0.15 Hz and 0.4 Hz.

4 Types of workload

The studies in the related work section provided possible approaches to measure workload. Pickups conceptual framework of mental workload for railway signalers[5] has been used to provide an accurate and complete representation of the workload of rail signalers. The framework is depicted in figure7. Three types of load are defined, the imposed load depends on task characteristics and is the same for all operators. The internal and perceived load depend on individual characteristics of the operator. For the perceived load there is a feedback loop from the work result. Results will influence the mental state of an operator and will effect his perceived load. Goals and strategies influence the physical and cognitive demand of operators and therefore the effort of operators. This will effect the performance and wellbeing and indirectly the perceived, subjective workload.

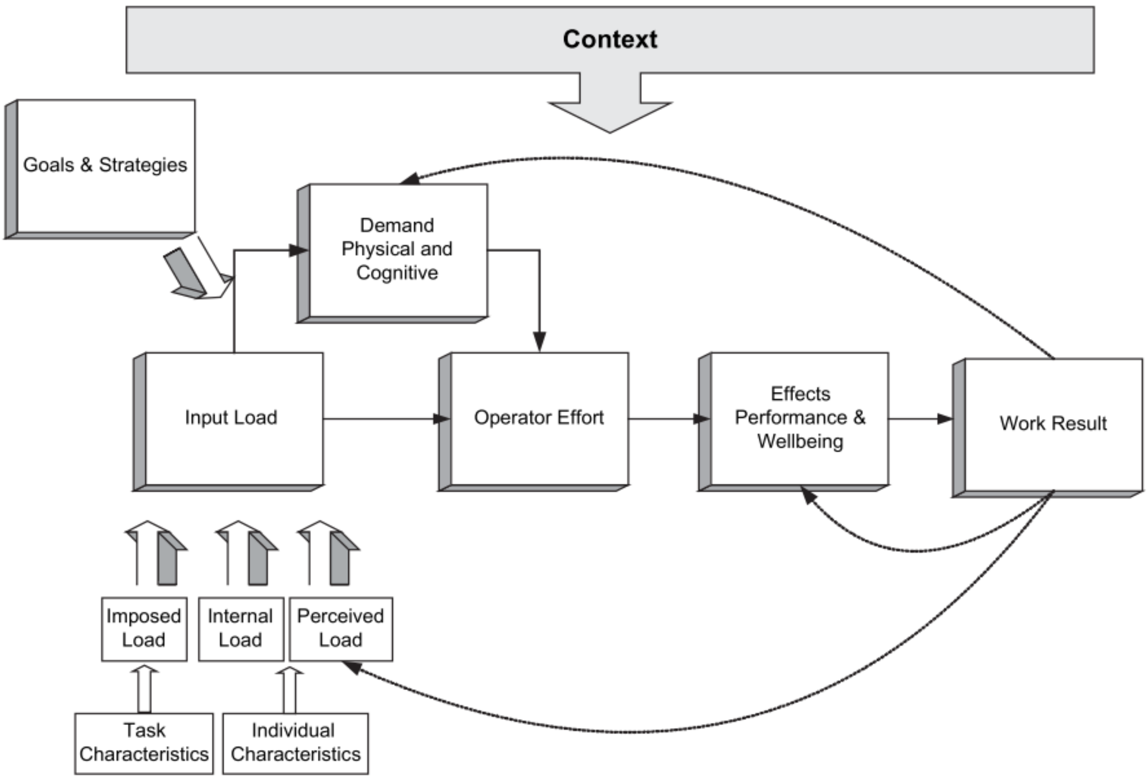

Figure 7: Mental workload framework by Pickup[16]

The imposed, internal and perceived load are the inputs for measuring workload according to Pickup's framework. Applying this framework to measure workload at a ProRail control post has be done by mapping these loads to appropriate workload types. The imposed load is based on task characteristics and has been applied at ProRail by measuring the performed actions of operators and their cognitive demands, Neerincx'[2] model for cognitive workload has been used after adapting it to the situation at the ProRail control post. The internal and perceived load are based on characteristics of the operator. Pickups[6] method to measure subjective workload using the IWS scale and the widely discussed HRV to measure physiological arousal have been used for the internal and perceived load. Veltman[17] proposed similarly that performance, subjective and physiological data are required for a complete understanding of workload. This resulted in three types of workload that were used for measuring the workload at ProRail control post Zwolle:

1. Cognitive workload
2. Subjective workload
3. Physiological workload

Combining these three types of workload has not been done and has not been applied to the railway sector before.

4.1 Cognitive workload

Neerincx[2] described his model for cognitive task load consisting of task complexity, task switches and task duration. These three components have been used as a basis to define a cognitive workload in the context of rail signalers. Task complexity has been determined by listing the tasks of rail signalers and categorizing these tasks according to their complexity. For task switches there needed to be determined how many tasks are handled by rail signalers and how many (external) events are of influence on their performance and therefore workload. Task duration has been determined by looking at the number of tasks for each interval of 5 minutes. A 5 minute interval has been used for the other types of workload (subjective and physiological) as well. Using the same interval made it possible to lists the scores for the different types of workload for the same interval.

4.2 Subjective load

Pickup's IWS scale[6] has been used to measure the subjective workload. It had already been used in practice by Wilms and Zeilstra[9], but they were unable to find a correlation between objective (cognitive) workload and subjective workload. A key aspect was that they measured IWS for periods of one hour, in this approach the duration has been extended by measuring IWS scores for longer periods. Scores were also collected manually in Wilms and Zeilstra's study, a researcher needed to be present for the complete duration of the experiment. Researcher have been omitted by automating the process of gathering IWS scores, a potential benefit for this approach was the ability to up-scale the experiment (measuring subjective workload at multiple workstations simultaneously) and making it easier to measure subjective workload for longer periods of time.

4.3 Physiological arousal created by workload

Heart rate variability (HRV) is listed as an appropriate way to measure physiological arousal and therefore to determine the physiological workload[10][12]. The heart rate of rail signalers needed to be recorded in order to calculate HRV. Each rail signaler needed to wear a heart rate sensor during their work. The HRV was calculated in the time domain[15] and has been used to determine the physiological workload.

5 Applying methods at ProRail

The cognitive, subjective and physiological workload were applied at ProRail post Zwolle for measuring workload. These methods were based on existing methods used in studies handled in the related work section which focused on measuring workload in a control room setting. The ProRail post is a control room setting as well, but it can nonetheless be different from the settings used in these studies. Another factor that was different for these studies was that the measurements at ProRail continued for 24 hours a day and the measurements took place in an operational setting, which was quite different from test settings used in some of the other studies. The requirements listed in section 1.3 were used to check whether the selected methods were applicable for usage at ProRail and to check how the methods needed to be changed to be applicable.

5.1 Checking requirements

Requirements **R1** till **R5** were checked for the following methods:

- Cognitive task load by Neerincx[2] (Cognitive workload)
- Integrated Work Scale by Pickup[5] (Subjective workload)
- Heart rate variability[11][12] (Physiological workload)

The checked methods were selected from the related work section and could be used to measure cognitive, subjective and physiological workload. In order to use these methods at ProRail there was first checked if the requirements, used to determine if a method is suitable to use in a real-time setting, were met.

R1 *No interference with normal routine of rail signalers*

The cognitive task load by Neerincx met the first requirement, since the input for this model could be obtained without interference of the rail signalers. The IWS by Pickup was used by asking the rail signalers in person to rate their workload. **R1** was not met since this involved direct interference with rail signalers. This requirement was also not met for the HRV, used to determine the physiological workload. Measuring the heart rate of rail signalers is required to determine the HRV, implying that rail signalers should wear a heart rate sensor which interferes with their routine.

R2 *Minimize required input from rail signalers (no more than one action per five minutes)*

Requirement was met for the cognitive task load, no input was required. It was also met for the IWS, only one action per five minutes was required. No input from rail signalers was required to determine the HRV, **R2** was therefore already met.

R3 *Limit set up time as much as possible (within five minutes)*

There was no setup time for the cognitive task load since no software or other device were needed at the time of measurements. The requirements was also met for the IWS. Only scores were collected, no set up time was required. A heart rate sensor was required for the rail signalers to determine the HRV. This required more time to set up, so the requirements was not met.

R4 *Keep systems running, no input from the researches should be required once the system has been set up*
Input from researchers was not required for the cognitive task load model, **R4** was therefore met. Researchers needed to provide input for the IWS method on the contrary and always needed to be present during the measurements, the requirement was therefore not met. For HRV there was no input required once the heart rate sensor was in place, the requirement was therefore met.

R5 *Keep systems running for longer periods, ideally 24/7*
Cognitive task load was determined afterwards, so it was already possible to measure it over longer periods. A researcher needed to be on the spot to measure IWS scores, it was therefore not possible to measure IWS scores for longer periods. There are multiple work stations at the ProRail post, so it was not possible to measure the subjective workload on all work stations simultaneously. Rail signalers needed to wear a heart sensor to measure the physiological workload. It was not desirable to let rail signalers wear the heart rate sensor for long periods, since it is not comfortable to wear it.

Table 1 provides an overview of the requirements for each method.

5.1.1 Cognitive task load

The cognitive task load as described by Neerincx[2] meets the criteria for measuring workload at the ProRail control post. The problem was that it is a theoretical model. The workload of operators needed to be determined by taken the task complexity, task switches and the duration of tasks into account. This information could be gathered from logging, so there was no need to disturb operators or interfere with their tasks. Applying this method at the ProRail control post could be done therefore. The challenge was to extract relevant data from ProRail logging. Neerincx' method needed to be adjusted since the data form ProRail was very specific and custom for their systems, Neerincx' method could be used as a guidance for determining the cognitive workload based on available logging data at ProRail.

5.1.2 Integrated Work Scale

For the IWS there were some problems with the approaches used by Pickup[6] and by Wilms and Zeilstra[9]. The main problem was that these studies focused on short research intervals and were therefore not applicable to be used in a real-time setting. **R1** was not met since there was interference with the rail signalers throughout the experiment, they were asked by a researchers to rate their workload at 5 minute intervals. The length of the measurement sessions was therefore limited. **R4** and **R5** were not met since input from researchers was necessary throughout the research, they had to ask for input continuously. Since they needed to be present throughout the experiment it was not necessary to keep the systems running for

	Cognitive task load	Integrated Work Scale	Heart rate variability
R1	✓	✗	✗
R2	✓	✓	✓
R3	✓	✓	✗
R4	✓	✗	✓
R5	✓	✗	✗

Table 1: Requirements compatibility for methods

long periods, let alone for 24 hours a day.

The method in itself was usable and relatively easy to implement. To meet the requirements it was necessary to limit the role of the researcher as much as possible. Having a researcher standing by cannot be done for long periods and has a large impact on the routine of rail signalers which was unacceptable during a measurement session in an operational environment.

5.1.3 Heart rate variability

Measuring the heart rate of rail signalers had a huge impact as well. Part of this impact was inevitable since a heart rate monitor will be required for heart rate measurements which are needed to determine the heart rate variability. To minimize the disturbance it was necessary to have comfortable heart rate sensors that required minimal maintenance, apart from the initial set up.

6 Methodology for measuring workload

The cognitive, subjective and physiological workload are the types of workload that were measured at the ProRail control post. The methodology for the measurements sessions describes how these types of workload were measured. The methods for measuring these types of workload were based on existing methods, that were adapted to be usable at ProRail. The methods had to meet requirements **R1** till **R5** as much as possible in order to be applied at the ProRail control post.

6.1 Cognitive workload

Cognitive workload was based on Neerincx' model[2] for cognitive task load. He defined the cognitive task load in the form of task complexity, switches and duration. ProRail data was used to map these aspects and to measure the cognitive workload of rail signalers.

6.1.1 Task complexity

Activities performed by rail signalers were used for task complexity, these consist of the actions that make up most of the work of rail signalers. There was distinguished between four activities:

1. Monitoring (Mon)
2. Planning (Plan)
3. Manual (Man)
4. Communication (Com)

Monitoring actions are activities performed by an automated system, these consist of setting train paths. This is setting switches to the correct position and setting signals in such a way that a train can enter and leave a station via the correct tracks. There is a automatic train path setter, in Dutch ARI (*automatische rijweg insteller*), that sets train paths based on a predefined plan. The plan lists the entry and exit time of a train for a specific time and station. When trains run according to schedule the plan can be executed automatically, but in case of delays or other calamities it can be necessary to adjust the plan. Examples of such plan mutations are adjusting arrival times or entry and exit tracks for a specific train. Plan mutations are performed by rail signalers and formed the second kind of action. Manual actions are operations on the infrastructure such as setting train paths, but now performed manually instead of automatically by ARI. Other manual actions are maintenance instructions, in Dutch WBI (*werplek beveiligings instructie*), and acknowledging (in Dutch *kwiteren*) of broken overpasses or other track elements. Communication actions are phone conversations between rail signalers and other parties like train drivers or the police.

The combination of these four activities formed the task complexity as described by Neerincx. The difference was that Neerincx used the Skill-Rule-Knowledge model by Rasmussen[3] to rate each task individually. This approach was not practical to use at ProRail since the described activities form relatively small actions and there can appear dozens of such actions in an hour. The tasks complexity was therefore not determined for each task but for each activity throughout the day.

6.1.2 Task switches

Task switches were derived from the number of task activations that were going on at a particular moment. Task activations are events that are taking place and are of importance to rail signalers. The number of task activations at a specific interval were taken into account and were used to fulfill the role of task switches in Neerincx' model. The following task activations were taken into account:

1. Delayed trains
2. Phone calls
3. Disruptions
4. Maintenance work

Delayed trains are counted when the delay exceeds three minutes (this is the definition used by ProRail) within the area controlled by the rail signaler. Phone calls is the number of phone calls made by a rail signaler. Disruptions are the number of defective signals, switches or overpasses. Maintenance are the number of maintenance activities on rail sections within the area of the rail signaler.

6.1.3 Calculating the cognitive workload

Data for both the activities and task activations was retrieved from ProRail logging data. Several logs from different systems are available at ProRail. One of the most important data sources was processing data that is logged for each work station, in Dutch PRL (*processleiding*) data. These process logs contain data for work stations and was used to retrieve monitoring actions, manual actions, disruptions and maintenance events. Plan data is also included in the process data, but is not logged per work station, this is logged per emplacement (station) instead. Planning actions can be linked to a work station since it is known which emplacement is controlled by which work station. Delayed trains were retrieved from activity logs, information about trains is also logged per emplacement and could be linked to a work station using this information. Phone logs are kept separately and were available per work station, these logs could be used to retrieve communication actions and the number of phone calls.

Each of the activity actions was counted in 5 minute time slots, which was used in the measurement of other workload types as well. Monitoring, planning and manual actions were counted for each interval and the values were normalized in order to be able to compare them. Normalizing the values was done for each of this three actions by dividing the value by the maximum of that particular action during the research period. Normalizing resulted in values between 0 and 1. Communication was determined by the percentage of a 5 minute time slot that was dedicated to phone calls, this resulted in a value between 0 and 1 for each time slot as well. This resulted in lists of normalized values for the activities (monitoring, planning, manual and communication) that represent the task complexity.

Task switches were represented by the number of task activations that were also counted for each 5 minute time slot. These values were normalized as well by dividing the number of task activations in a 5 minute time slot by the maximum number of task activations that occurred in a 5 minute time slot during the research period. This resulted in values between 0 and 1 for task activations for each time slot.

The third aspects of Neerincx' model, the task duration, was represented by the fact that the cognitive workload was determined in time slots of 5 minutes. The overall result for cognitive workload was a score for each 5 minute time slot.

Calculating the score for the cognitive workload was done by combining the values for activities and task activations into a single score for each 5 minute time slot. The cognitive workload as calculated here is different from Neerincx' cognitive task load. Neerincx' provided a theoretical model for the cognitive task load. The aspects of this model (task complexity, duration and duration) were used as a basis, but used in the context applicable for ProRail, using available logging data. This newly developed measurement is called the external cognitive task load (XTL), since the task activations are external events that influence the cognitive state of rail signalers and to distinguish from Neerincx' cognitive task load. The first step in calculating the XTL was adding the normalized activity values. Neerincx used the Skill-Rule-Knowledge model to determine the task complexity, in this case the activities throughout the day. The cognitive relation between the activities was unknown, the complexity was therefore added to the scores using a different approach. This was done by multiplying each activity with a constant, respectively; K_{mon}, K_{plan}, K_{man} and K_{com}, to distinguish between the complexity of the individual activities. The values for these constants were initially set to 1 and could be changed by empirical investigation.

The next step was to include task activations, the external factor. This represented the task switches and should act as an amplifier for the XTL score. The number of task activations (switches) for a time slot were taken into account, effectively there can be 0 till n task activations. They should act as an amplifier since more task activations require more attention from rail signalers. A switching factor (K_{switch}) that amplifies the scores in such a way that 0 switches have no influence on the XTL score and the maximum numer of switches doubles the XTL score was introduced to realize this. The value of the switching factor lies between 1 and 2. The switching factor was calculated in the following way:

$$K_{switch} = \frac{number\ of\ task\ activations\ in\ 5\ min\ time\ slot}{max\ number\ of\ task\ activations\ in\ 5\ min\ time\ slot} + 1 \quad (1)$$

The formula used to calculate XTL was the switching factor times the normalized values for activities multiplied with their respective constants. The formula for the XTL score for a single work station for one time slot was:

$$XTL_{WS} = K_{switch} \times (K_{\text{mon}} \times \frac{Mon}{Mon_{max}} + K_{\text{plan}} \times \frac{Plan}{Plan_{max}} + K_{\text{man}} \times \frac{Man}{Man_{max}} + K_{\text{com}} \times Com) \quad (2)$$

The XTL scores for the target (G_T) and reference (G_R) group were calculated for each time slot by averaging the sum of the XTL scores for the work stations in the corresponding group.

6.1.4 Processing log data

The data to calculate the XTL scores was extracted from ProRail log data. A data handler was implemented to process this log data and to calculate XTL scores automatically. The handler consisted of an import side that processed the required log files and stored the relevant information in a database. Relevant log data is stored in a database without any transformations. An export module was used to calculate XTL scores per day for all work stations.

6.2 Subjective workload

Subjective workload was measured using the Integrated Work Scale (IWS) developed by Pickup[5]. The IWS was used by rail signalers to rate their subjective workload on a scale from 1 till 9. Scores were gathered manually by Pickup by simply asking the rail signalers for their ratings at fixed intervals. This approach had a number of disadvantages, the most important one was

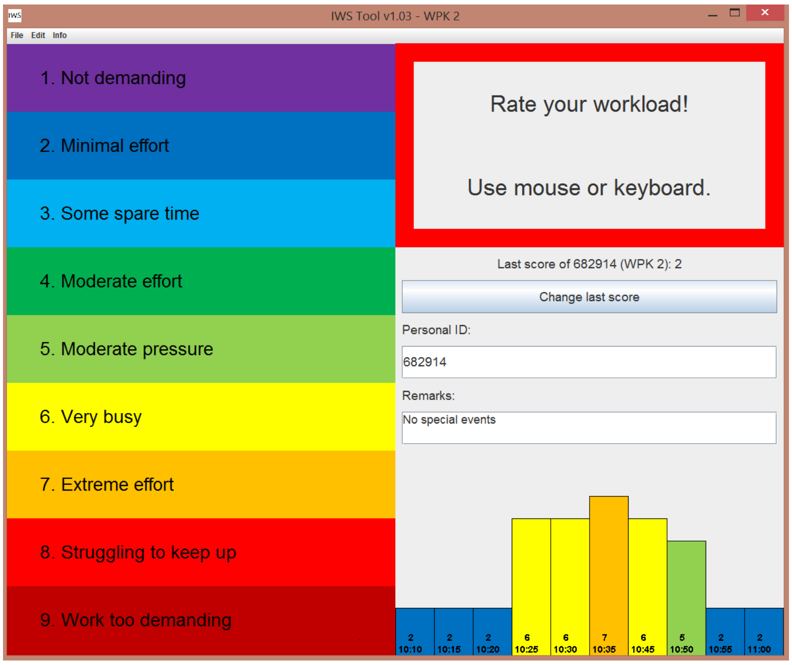

Figure 8: IWS interface for rail signalers

that rail signalers were continuously bothered during their work (**R1**). Gathering IWS scores for longer periods was not possible using this approach (**R4** and **R5**). The process of gathering IWS scores was automated to overcome these issues.

6.2.1 IWS user interface

A user interface for rail signalers was developed that could be used to gather IWS scores throughout the day. The interface with the IWS scale was available at the work station of each rail signaler. Scores needed to be entered at a five minute interval, a warning 'bleep' and a red blinking rectangle in the interface reminded a rail signaler to enter a score. The score represented their subjective workload on a scale from 1 to 9. The interface was kept as simple as possible, since the rail signalers should not be distracted. The left half of the screen contained the IWS scores. Rail signalers could click on the score or press the corresponding number on the keyboard to register their score. The right half of the screen was split in three blocks, the upper block contained a timer that indicated when to enter the next score. Rail signalers could change their last entry, provide their personal ID (only necessary at the beginning of a session) and could leave comments in the middle block. A graphical representation of the last ten scores including time stamps was shown in the lowest block.

See figure 8 for the IWS interface (terms were in Dutch in the version presented to the rail signalers). Scores had to be entered at 5 minute intervals, when a rail signaler forgot to enter a score it was filled backwards, meaning that the new score was entered for the previous missed intervals. There was also the possibility to correct the last entered score and to leave comments. Scores were stored in log files and were imported into the same database used for XTL data.

6.2.2 Monitoring interface

A monitoring interface was available for observers (the researchers), see figure 9. The interface provided an overview of all work stations by showing graphs of the latests scores. Data was gathered real-time from the log files of each work station. The last 12 entries for each work

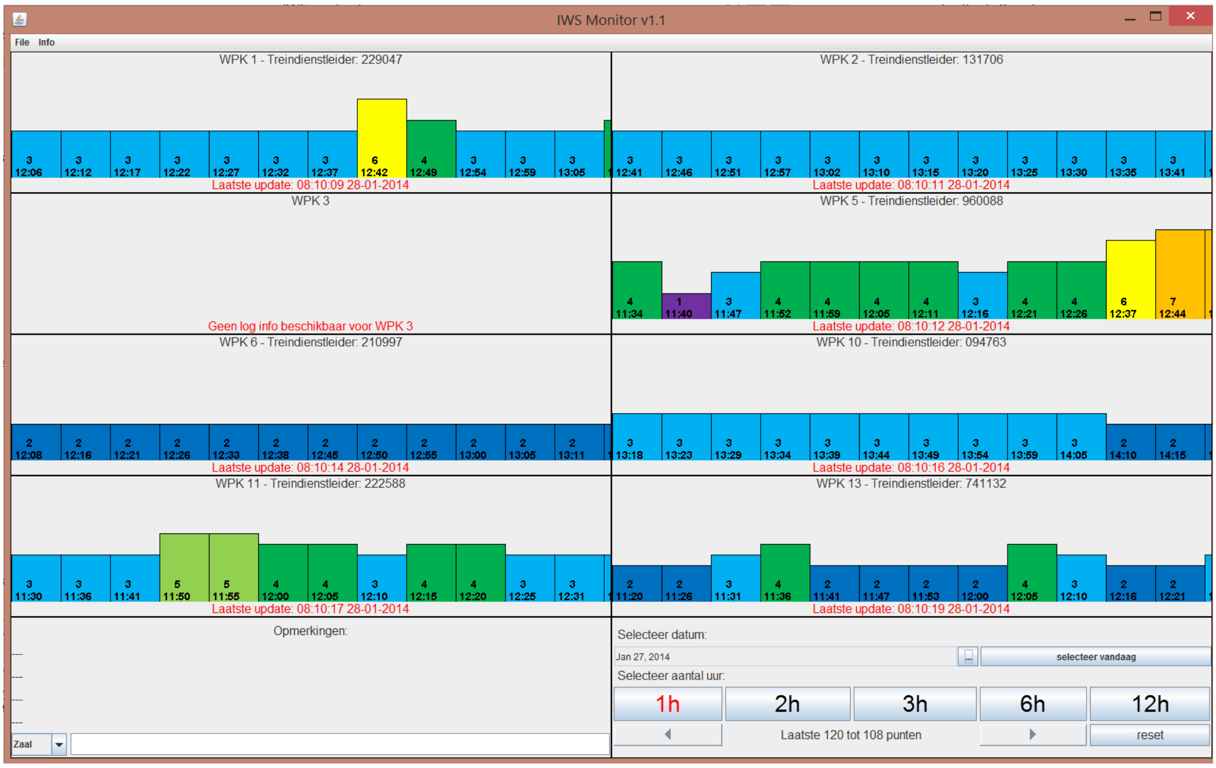

Figure 9: Monitoring interface

station were shown, this corresponded with the last hour since a 5 minute interval was used. The view could be adjusted to show more data and it was possible to browse through the history of IWS scores for each work station. Main advantage of this overview was that it was easy to spot busy work stations, before it was required to check each work station individually.

6.3 Heart rate sensor

Physiological workload was measured using a heart rate sensor. The Zephyr HxM BT heart rate sensor was used, this sensor can be strapped to a belt and worn on your chest. Heart rate data was gathered via blue-tooth and was logged on a nearby laptop. The sensor transmits the 15 latests R-R intervals, a time stamp and some meta data every second. An advantage of this particular sensor was that it sends the latest 15 R-R intervals each time it transmits. Gaps in the data could easily be restored as long as the signal was picked up at least every 15 seconds by the laptop.

R-R intervals were collected for 5 minute time slots to be in line with XTL and IWS scores. The R-R intervals were used to calculate the average heart rate in a time slot, but this is not useful by itself. The heart rate variability (HRV) was much more interesting in terms of the physiological workload of rail signalers. Togo and Takahashi[15] described methods to determine the HRV, the most common method was calculating HRV as the standard deviation of all normal-to-normal intervals (SDNN). The HRV score was determined for each time slot by taking the SDNN of all R-R values in this time slot. Erroneous R-R values were filtered from the results before calculating the HRV. Errors occurred sometimes when the sensor was unable to measure the heart rate, these erroneous values were obviously wrong since they were way too high or low. Most errors occurred when strapping the sensor around the chest or when the belt was taken of.

Log files containing the average heart rate and SDNN of R-R values for each time slot were stored for each work station in the same database as before in order to match the HRV data to XTL and IWS data later on.

6.4 Stretches

Scores for XTL, IWS and HRV were available for each five minute time slot for each work station. The scores already matched in time since the same interval was used for each type of workload. The different types of workload could be compared for these intervals. IWS scores, form the subjective workload and were validated through the HRV. Changes in the IWS scores were noticeable in the HRV as well since a physiological reaction occurred when rail signalers rated the subjective workload at a high level. XTL and IWS could be compared for each interval, but these are relatively short periods since time slots of five minutes were used. It makes more sense to compare these types of workload for a longer period, since events that are worthwhile to compare have a longer duration as well. Events could be disruptions or other calamities that were of influence to rail signalers. Such events can not be found as single entries in the logging data. Events are typically a cluster of task activities so they were gathered indirectly from the logging data based on XTL scores. The steady state for a rail signaler is when all train activities are running according to plan. Events were defined from the moment of a disruption, when the steady state was broken, to the next occurrence of the steady state. Multiple events could be captured in such periods since a steady state is used as a base-line.

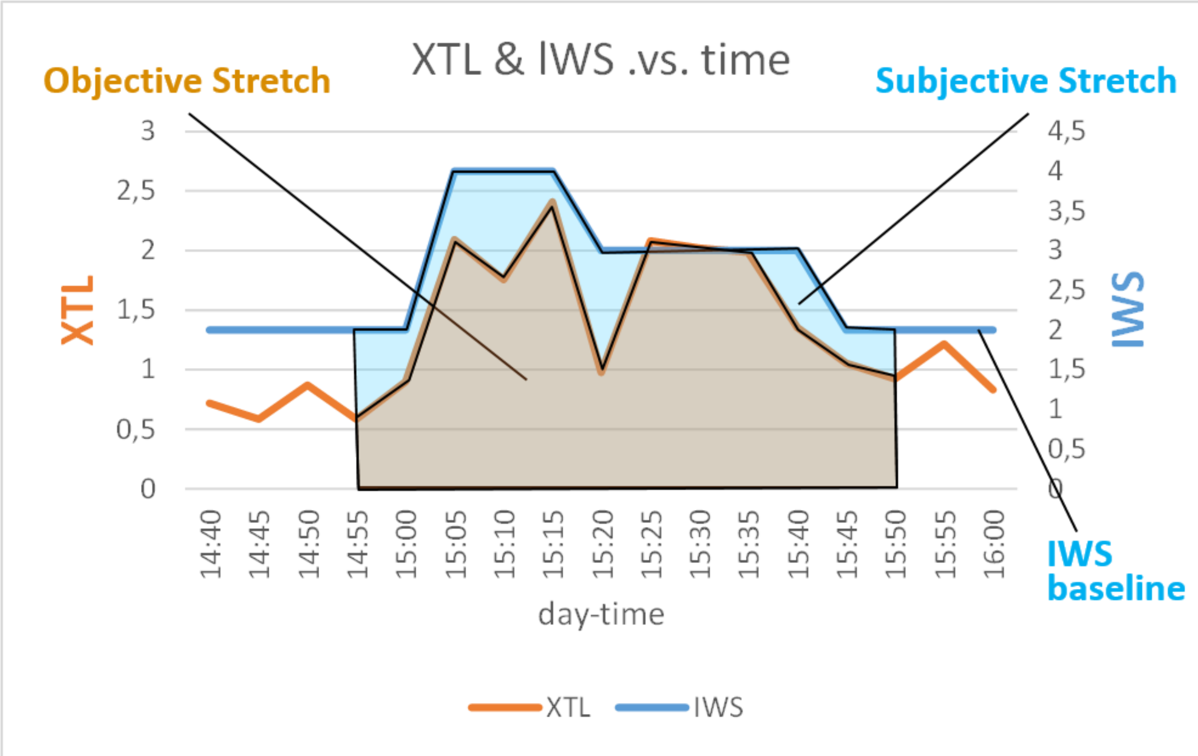

Figure 10: Objective- and subjective stretch of XTL and IWS in time

A new metric called 'stretch' was introduced to compare XTL and IWS values during these events. The IWS base-line was used to define the steady state of the system. The start of a stretch is when the IWS score deviates from its base-line and the end of the stretch is when the IWS score is at its base-line again. IWS scores were a reaction of rail signalers on system activities, the real start and end of a stretch were therefore not always in line with the IWS scores. The start of the stretch was therefore adjusted to the first minimum of the XTL score before the rising of the IWS score and the end was adjusted to the first minimum of the XTL score after the IWS score reaches its base-line again. See figure 10 for a visual representation of a stretch.

A stretch depicted a reaction of a rail signaler, because of its deviation from the IWS base-line, to a (cluster of) events. The objective stretch is the area in the graph of a stretch under the XTL, because the XTL score represents objective measurements. The subjective stretch is the area under IWS, since this represents the subjective rating of a rail signaler. Finally the ratio between subjective and objective stretch (subjective divided by objective stretch) was defined as the stretch-ratio. The stretch-ratio could be be used to compare two periods of events (two stretches) with each other. The stretch-ratio was useful since it allowed comparing two stretches of different lengths. Stretches with a similar stretch-ratio are an indication that the balance between subjective and objective workload is similar. When the stretch-ratio was above average for a certain period it was an indication of a higher subjective rating of external events that were lower rated on average. Comparing a large amount of stretch-ratios, for instance all stretch-ratios of stretches for a day, was useful in comparing the relative workload between these periods. If the average stretch-ratios for the days were different is was an indication for a change in the workload.

7 Measurement sessions at ProRail post Zwolle

Two measurement sessions were held at ProRail post Zwolle. The sessions surrounded the reorganization at post Zwolle in December 2013. The first session, in which both the target (G_T) and reference (G_R) group had the same tasks for rail signalers, was directly before the reorganization. The measurements were held from Monday November 25 till Friday November 29, 2013. The measurements took place on a daily basis from 7:00 AM till 9:00 PM. The second session was two months after the reorganization, this was chosen to eliminate effects from getting used to the new situation. It was held from Monday January 27 till Friday January 31, 2014. The measurements were extended to 24 hours per day during this second session, so night shifts were included as well.

7.1 Differences

The main difference was that G_T worked according to the reorganization. Tasks for rail signalers were split in safety and logistic tasks. First there were three active rail signalers in G_T and one extra rail signaler who could assist in rush hours or in case of calamities. After the reorganization there were three rail signalers for safety tasks and one rail signaler for logistic (planning) tasks for the complete group. G_R worked similar to the situation before the reorganization.

There were a few changes between the two measurement sessions. The night shift was included during the second week and the IWS interface that rail signalers used to rate their subjective workload was on a laptop at their work station during the first week, but was integrated in their operational system during the second week, see figure 11.

7.2 Research population

The population for the research consisted of 44 participating rail signalers, 79.5 % of the group were males. Ages varied from 23 to 64 years, with a mean of 43.6 years. Experience as rail signaler varied between 0 and 37 years, with 17.6 years on average. XTL data is gathered from ProRail logging and 100 % of the rail signalers took part in the IWS measurements. There is therefore a complete data set of all work stations for XTL and IWS data. Only 39 % of the rail signalers took part in the HRV measurements. The main reason for not participating with the HRV measurements were objections against wearing a belt with the heart rate sensor.

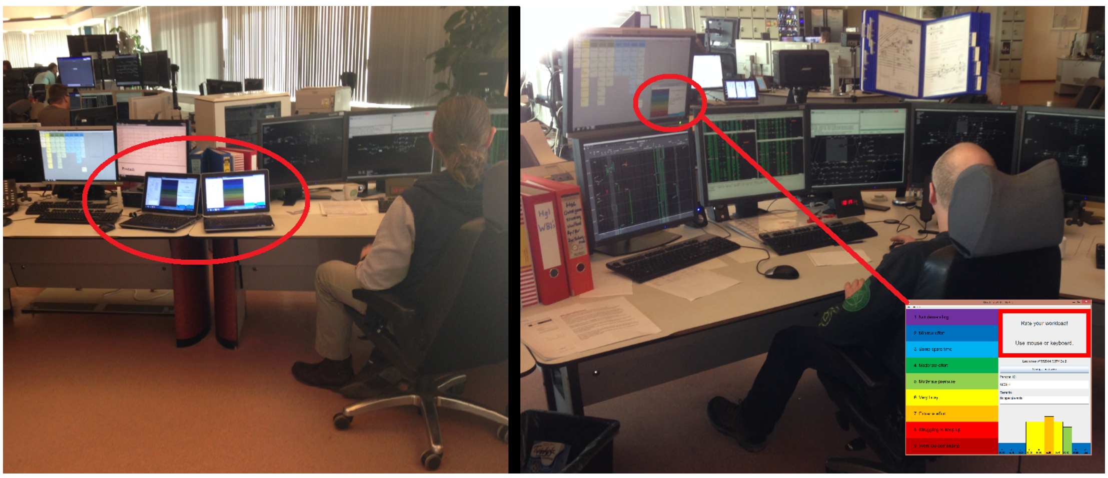

Figure 11: Left: IWS via laptop (1st session). Right: Integrated IWS (2nd session).

8 Results

Results consist of the integration of workload measurements in the operational environment. Measurements were integrated in a real-time setting at ProRail. The second part of the results is derived form data gathered during the measurement sessions.

8.1 Integration in operational environment

The requirements from section 1.3 were used to check if the workload measurements were successfully integrated in an operational environment with minimal inconvenience. Three types of workload have been measured, the external cognitive task load (XTL), the subjective load (IWS) and the physiological workload. The XTL could be derived from ProRail logging data (process and phone logs) and was determined afterwards and caused no inconvenience for the rail signalers. The IWS measurements required input from rail signalers at five minute intervals. Input was limited, rail signalers rated their workload on a scale from 1 to 9 by pressing a single key or clicking a button. During the first week there was a separate laptop with the IWS interface for each workstation to rate the workload. The IWS interface was integrated in the operational systems for rail signalers during the second week. Integration in the operational system was preferred since it made the interface available on a screen that rail signalers were already monitoring.

This new approach for gathering IWS scores was an improvement compared to Pickup's approach were rail signalers were continuously asked to rate their workload. **R1** (no interference with routine of rail signalers) was not completely covered, there was still interference required since rail signalers needed to enter their scores. **R2** (minimize amount of input to one action per five minutes) was covered, input was inevitable since IWS scores needed to be gathered, but it was minimized to one action per five minutes. **R3** (limiting set up time) was covered, set up time was limited especially during the second measurement week when the IWS interface was integrated in the operational systems. The only thing that needed to be entered was the personal ID of the rail signaler at the start of his shift. **R4** (no input from research during measurements) was covered as well, researchers were no longer required once the system was running. **R5** (run measurements for longer periods, ideally 24/7) was also covered, the IWS interface ran 24/7 without any problems as was demonstrated during the second measurement session in which the IWS interface was running for a full week.

A heart rate sensor was used to measure the physiological workload. The sensors needed to be strapped on a belt that was worn on the chest. This caused some inconvenience for the rail signalers especially when they needed to put the belt around their chest. This inconvenience resulted in a low participation score of 39 % for the heart rate sensor. Not all requiremenrs were met for the physiological measurements. **R1** was covered since the measurements do not directly interfere with the routine of rail signalers, but wearing a heart rate sensor does cause inconvenience for some people. **R2** was covered, no input from rail signalers was required. **R3** was covered since the set up time was limited, the only required action for the rail signaler was putting the belt with the heart rate sensor on. **R4** was not covered, the heart rate sensor sends its data via a blue-tooth connection to a nearby laptop. This connection failed when a rail signaler left his work station, to get some coffee or went to the bathroom for instance. The heart rate sensor was not able to reestablish its connection once the laptop was in reach again. This was solved by the researchers by manually restarting the application. **R5** was not met as well. It was possible to gather data from the heart rate sensor for long periods, but most rail signalers were not comfortable wearing the sensor for long periods.

	Group	## Stretch	Stretch-ratio		Subjective Stretch (SS)				Objective Stretch (OS)			
			Mean [IWS/XTL]	SD	Mean [IWS x min]	SD	Mean(SS/Dt) [IWS]	SD(SS/Dt)	Mean [XTL x min]	SD	Mean(OS/Dt) [XTL]	SD(OS/Dt)
Before reorganization	Target	35	5.30	2.61								
	Reference	107	5.82	2.55								
	All (T & R)	142	5.69	2.57	21.13	15.60	3.09	0.80	4.28	3.58	0.62	0.26
During reorganization	Target	170	7.37	4.24								
	Target - planner	134	6.17	2.81								
	Reference	134	6.36	1.80								
	All (T & R)	304	6.92	3.42	21.17	24.30	2.75	0.59	3.49	3.82	0.47	0.21
	All - planner	268	6.26	2.36	21.18	25.59	2.75	0.59	3.70	4.00	0.50	0.20

Figure 12: Overview of stretch measurements

8.2 Stretch measurements

Stretches were tagged by hand for each work station according to the rules explained in section 6.4. Tagging of stretches was done for all work stations for every day of the two measurement sessions. Figure 12 provides an overview of the measurement results derived from stretches. For the first week the mean of the stretch-ratio for the target group was 5.30 with a standard deviation (SD) of 2.61. The mean stretch-ratio for the reference group was 5.82 and the corresponding SD was 2.55. The SD was relatively high for both groups, so there was variety in the stretch-ratio meaning that not all XTL scores were rated with the same IWS scores. The stretch-ratio mean and SD were in line for both groups, there can be concluded that the workload for the target and reference group was in the same order of magnitude.

The subjective stretch (SS) lists the mean in the first column. This is the average value of the sum of IWS scores for a single stretch. The SD in the second column is over this average value. The SD is quite high (15.60) meaning that the length of stretches was highly variable. More interesting is the Mean(SS/Dt) column, this is the average IWS score during a stretch, the fourth column, SD(SS/Dt), is the SD of this average IWS score. The columns for the objective stretch (OS) are similar. Interesting is that the average IWS score during a stretch is 3.09 and the average XTL score is 0.62. The length of the average stretch is about 35 minutes, this is calculated by dividing the mean of the SS by the average IWS score time 5 minutes (length of time slot). The same calculation can be made with these columns for OS.

The bottom half of the table is for stretches after the reorganization. Note that there are two rows for the target group, one for the overall scores and one for scores without the planner work station. The mean for [IWS/XTL] for the planner work station was 11.83 with a SD of 5.54 (numbers are not in this overview table). The planner deviates from the other work stations, the reason for this much higher stretch-ratio is that there is much less work on the planner work stations (no monitoring or manual actions and fewer communication) resulting in very low XTL scores, whereas the IWS scores are similar compared to other work stations. A row without planner data is therefore included. More details on the difference between the planner and other work station will be handled in section 8.4 about work distribution.

The mean stretch-ratio after the reorganization was 6.17 for the target group (without planner) and a SD of 2.81. For the reference group the mean stretch-ratio was 6.36 with a SD of 1.80. The workload was again similar in the target and reference group. The stretch-ratio values were higher, which was caused by fewer phone calls during the second week. Phone calls (communication actions) form and important factor, which is handled in section 8.4.2 on work distribution. The overall conclusion was that there were no significant differences in the workload distribution over the two measurement sessions.

8.3 Stretch validation by HRV

Subjective stretches were validated by HRV values, when a stretch occurred the subjective workload (IWS) score went up. This effect should be noticeable in the HRV values. The HRV should decrease in such a situation according to Togo and Tahakashi[15]. This effect has been analyzed for the first week. The highest value of HRV on the boundaries of a stretch was marked, multiplied by the stretch duration and subtracted by the integral under HRV for the complete stretch. A negative value confirms the subjective stretch according to the literature, this was the case in 83 % of stretches where HRV data was available.

8.4 Work distribution

The distribution of workload is an interesting aspect, especially for the target group. The target group was handled from three work stations before the reorganization, an extra planner work station was added after the reorganization. The effects of this extra work station are clearly visible in the distribution of activities for the target group, see figure 13. The distribution for the activities of rail signalers (monitoring, planning, manual action and communication) and task activations (events) is shown in this figure. For each activity there is shown for each work station in the target group how much activities occurred. The height of the bar is determined by the sum of all normalized actions for that specific activity during the week. A higher bar means a higher number of actions. The yellow line is the distribution between work stations for that activity. The distribution is the standard deviation over the work stations. For monitoring the distribution is quite high, since there is variability between the work stations. For communication the distribution is quite low, since communication activities occur to the same extent for the different work stations.

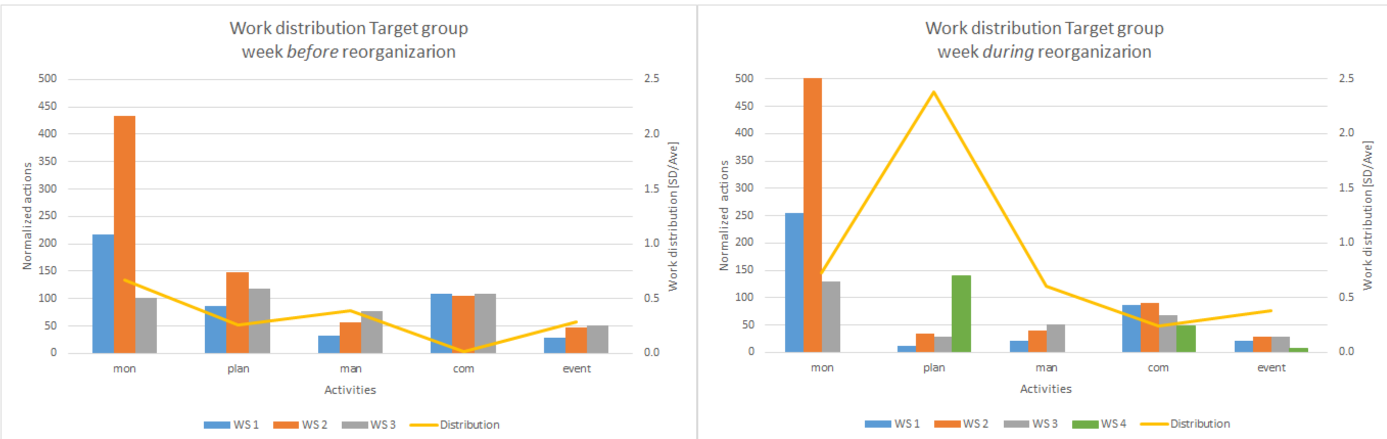

Figure 13: Work distribution target group

The planner work station (WS 4 on the right side in the figure) was added after the reorganization. The planner handled almost all planning actions after the reorganization and performed no monitoring or manual actions and had less communication compared to the other work stations. A note for planning actions is that due to limitations of ProRail logging it was not possible to log planning actions to an exact work station. Normally we were able to assign planning actions to a work station since the part of the infrastructure where the planning action was performed is known and this infrastructure could be linked to a work station. However, the planner handled planning actions of infrastructure that could be assigned to any of the three 'standard' work stations. In consultation with ProRail there was decided to assign plan actions for 10 minutes or more ahead to the planner work station and other plan actions to the corresponding 'standard' work stations. This seemed to be a correct approach based on observations.

There was also a large difference in monitoring actions, this is due to the infrastructure that is allocated to the work stations. Work station 2 handled far more monitoring actions both before and after the reorganization. Monitoring actions require relatively less effort from rail signalers since it are automated actions. Apart from planning actions the distribution was similar before and after the reorganization. The distribution of work for the reference group was also similar before and after the reorganization, see figure 14.

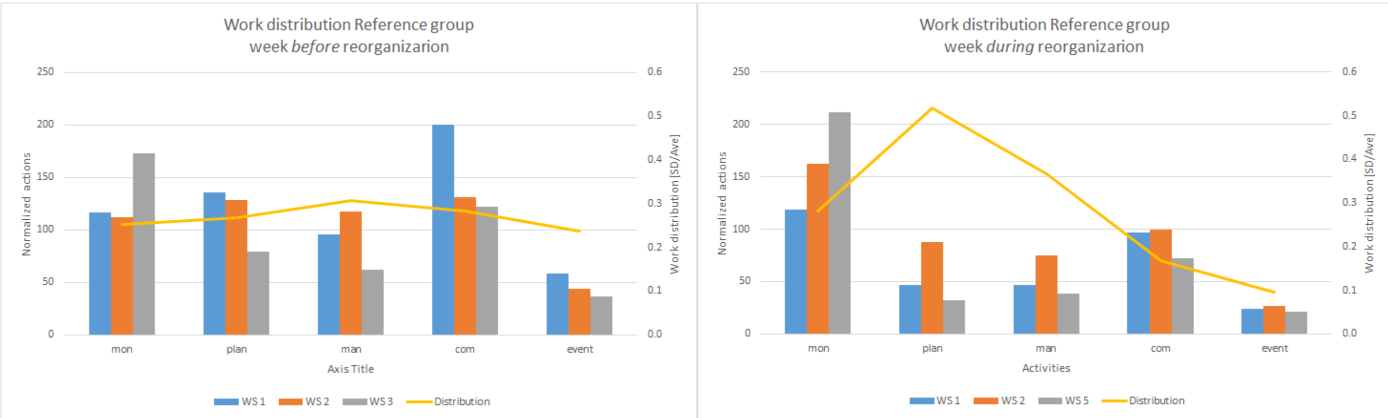

Figure 14: Work distribution reference group

8.4.1 Distribution IWS scores

The distribution of IWS scores for the second measurement week is shown in figure 15. From this distribution can be derived that the IWS scores were quite low, 62 % was a score of 2 (minimal effort) and only 0.6 % was a score of 6 or higher, which was considered for abnormal situations such as calamities. Rail signalers were experiencing underload for the majority of time. This can be explained by the fact that rail signalers only have to monitor when all train activities were handled according to plan. Rail signalers come into play when something goes wrong.

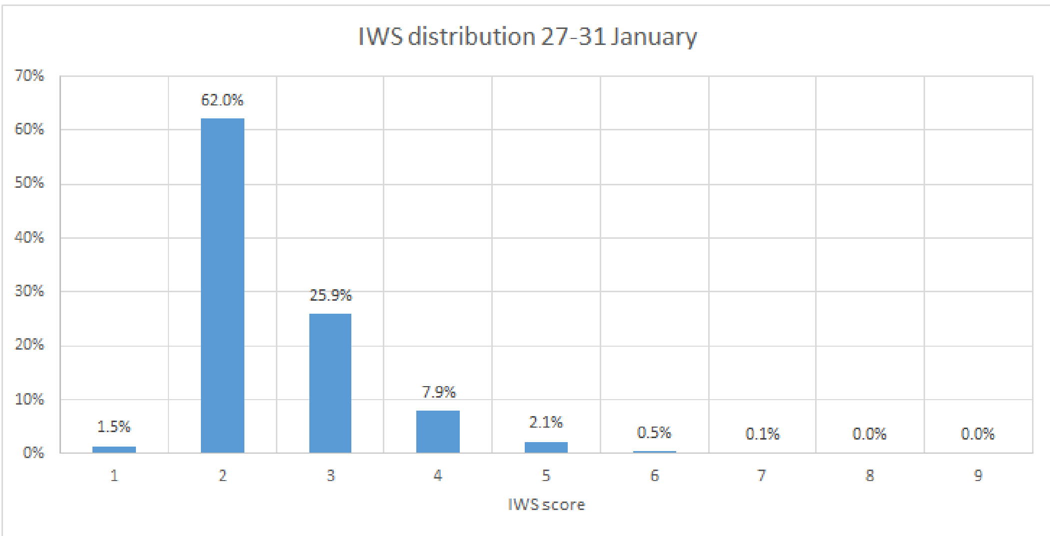

Figure 15: IWS distribution after reorganization

8.4.2 Distribution XTL components

Figure 16 shows the distribution of XTL components over the second measurement session. It shows the difference between the distribution of the normalized XTL components (monitoring-, planning-, manual activities and communication) over the complete session and for times when the subjective workload was high (IWS scores of 6 or higher). The orange bars show the average of all normalized values of the corresponding XTL components for each time slot and all work stations during the second measurement session. The blue bars were calculated similarly, but scores for XTL components were only used if the XTL score for that time slot had a score of 6 or higher. The values for each component can lie between 0 and 1 since the normalized values are shown.

The graph shows the difference in the distribution of the different XTL components, the orange bars represent the distribution of the components over the complete measurement session and the blue bars represent the distribution of the XTL components at the time of calamities. An increase is shown for all XTL components during calamities, but communication rises exceptionally. This shows that communication becomes a key factor during calamities and takes up a major part of the time of rail signalers. The average score of communication was 0.45 during calamities, this means that 45 % of the time was spend on phone calls in this situation.

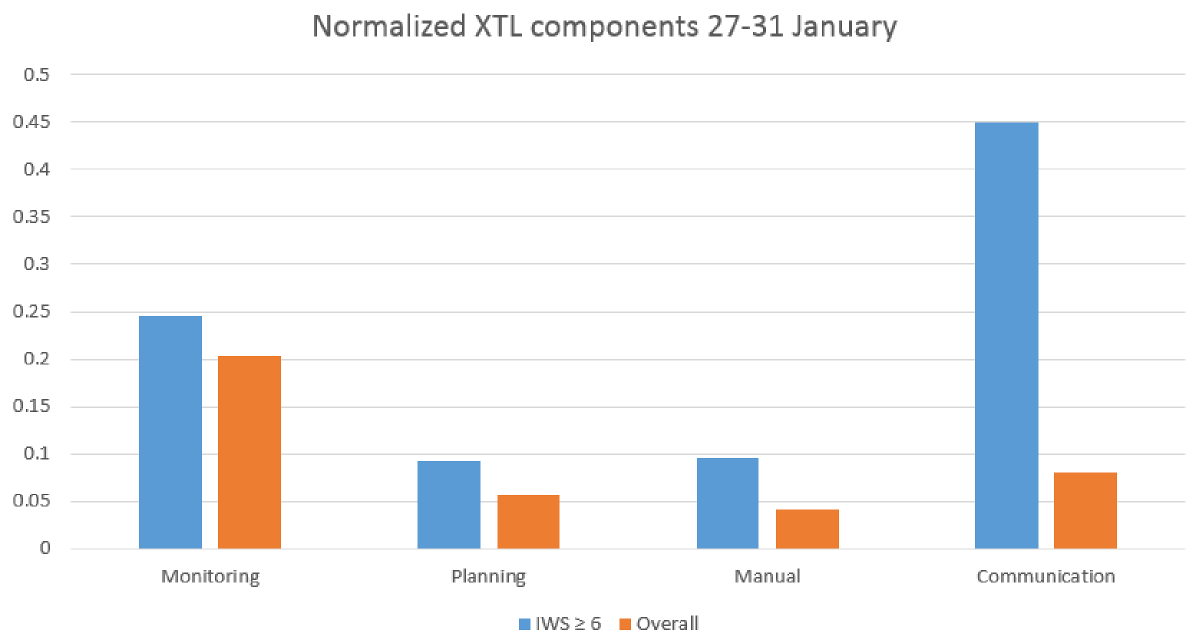

Figure 16: Distribution XTL components

9 Discussion

Measuring workload changes in an operational environment was the main goal for this research. Three measurements are used to measure workload: the external cognitive task load (XTL), the integrated work scale (IWS) and the heart rate variability (HRV). These measurements were based on existing methods and adapted to be used at a ProRail control post. XTL and IWS are combined in a new metric called 'stretch' which describes the objective and subjective changes in workload during a cluster of events. HRV measurements are used for validation purposes, which proved to be effective for validation of subjective stretches during the first measurement session.

9.1 External cognitive task load

XTL consists of four activities; monitoring-, planning-, manual actions and communication. The weight of these activities, defined by constants in the XTL formula, is set to 1 for calculating XTL scores. These constants can be set to put more weight on one of the activities, but this should be further investigated since the cognitive relation between the activities is currently unknown. Adjusting the constants will also result in a more equal distribution of the planner work station. The planner work station was kept from the results for now since the stretch-ratio was off due to differences in the XTL components.

Neerincx' methods for cognitive task load[2] was used as a basis for XTL. The main challenge was that Neerincx' model is theoretical and has not been applied before. ProRail logging data was used as input to calculate XTL scores. Most of the logging data was available within a few days and could be imported without modification by the data handler implemented during this research. Some parts of the data were more time consuming to handle. Phone data needed to be extracted manually by a ProRail employee. This resulted in multiple log files per work station per day. These log files needed to be merged manually which is an error prone approach.

The addition of the planner work station after the reorganization caused another difficulty since it is not possible to log plan activities directly to a work station as explained in section 8.4. Plan activities were allocated to the planner work station if it involved plan mutations for more than ten minutes in the future and were allocated to the 'standard' work station otherwise. This assumption was made in consult with a ProRail manager who knows how plan activities are executed. Observation during the measurement session showed that the assumption seemed to be correct. A complete guarantee of this approach cannot be given however, there are always exceptions which are now wrongly allocated. This problem can only be avoided if plan activities were logged per work station, but this has to be implemented at the side of ProRail.

9.2 Integrated Work Scale

The IWS developed by Pickup[5] was used to measure subjective workload. An interesting concept about the IWS interface is that it can be used by rail signalers to provide real-time feedback. When providing a high IWS score they are signaling that they are busy or are under stress. When the IWS scores are monitored it becomes instantly visible that something is going on. The IWS monitor provides an overview of the complete post in a glance.

The provided scores are subjective, IWS score may differentiate among the rail signalers. It was noticeable that most rail signalers had a IWS score base-line, that they used when they were just monitoring, scores of 2 or 3 were common. Rail signalers also copied their colleagues, there could be seen that rail signalers who were sitting next to each other used the same base-line for instance.

A disadvantage of subjective scoring is that rail signalers could be influenced by other things apart from their job when rating their workload. Personal problems could therefore influence the submitted IWS scores for instance. It is very hard to prove this effect, especially in what way scores are influenced by factors other than their job.

9.3 Heart rate variability

HRV was used to measure the physiological workload. The participation of rail signalers for the HRV measurements was quite low (39 %). Most rail signalers that refused to participate had objections against the heart rate sensors. Rail signalers did not like the idea of wearing a heart rate sensor or found it uncomfortable to wear the sensor around their chest. There are alternatives to measure the heart rate, for instance using a sensor around the wrist. Chances are however that the same objections still apply.

Another solution can be to use another measure to determine the physiological workload. Blink rate, pupil size or skin conduction can also be used as measurements. Changing the measurement leads to a completely different approach that might have other disadvantages such as the need for new measuring devices.

9.4 Overall results

The results over the two measurements sessions are constant. The average stretch-ratios for the target and reference group are in the same order of magnitude for both weeks. The reorganization had influence on the workload distribution for planning activities, which was expected. There was no detectable change in the overall workload however.

The two research sessions were too short to say something on the influence of the reorganization. Both measurement weeks were relatively quite weeks in which no serious calamities occurred. To put the workload measurements to the test is would be interesting to see what happens in such situations. It can therefore be useful to perform measurements over a longer period of time.

10 Conclusion

Measuring workload of a socio-technical system in real-time setting was the objective for this research. ProRail control post Zwolle was chosen as a setting since this control room is an example of a socio-technical system. Three types of workload were defined to measure at this setting; cognitive, subjective and physiological workload. Several studies for workload measurement exist, but they were not focused on a combination of these three types of workload and were not performed in a real-time setting.

The first step was to find existing methods for measuring these types of workload in a control room setting. Applying these methods to the ProRail control post was the next step. The workload of rail signalers at this control post was measured during two sessions of a week in which detecting changes in workload was the goal.

10.1 Existing methods to measure workload

The first goal was to find relevant methods to measure workload in a socio-technical setting. Cognitive, subjective and physiological workload were chosen as the types of workload to measure. Several studies for measuring these types of workload in control room settings have been performed in the past. Relevant methods for each of the workload types were investigated. The basis for cognitive workload is Neerincx' model on cognitive task load[2]. Task complexity, number of task switches and the duration of tasks are the key concepts for determining the cognitive task load. The Integrated Work Scale from Pickup[5] can be used to measure subjective workload. This method has already been used to measure subjective workload of rail signalers and was therefor suitable to use at the ProRail control post. The choice for physiological workload was the heart rate variability (HRV). This has been widely used for measuring workload[11][12]. The next step was to apply the found methods for cognitive, subjective and physiological workload to the setting at ProRail.

10.2 Applying methods for measuring workload to a ProRail control post

Adapting the methods to the ProRail setting was the next goal. Neerincx' cognitive task load model has been applied to available ProRail data which has led to the new measurement external cognitive task load (XTL). XTL represents the cognitive workload of rail signalers and is calculated on the basis of actions performed by rail signalers and task activations, such as delayed trains or disruptions, that are of influence to the rail signaler. All this relevant data is gathered from logging of ProRail systems.

The IWS used to measure subjective workload has been adapted for use at ProRail as well. The scale itself is left the same as Pickup, since it was already validated for measuring workload of rail signalers. An interface has been created so that rail signalers were able to rate their subjective workload on this scale without any further disturbance from researchers. The IWS interface was integrated in the operational systems that are already in place for rail signalers.

HRV was measured using a heart rate sensor that was worn by rail signalers. Data from the sensor was transmitted via blue-tooth to a laptop nearby. This data was used later on to determine the HRV.

XTL, IWS and HRV could be applied to measure workload at the ProRail control post.

10.3 Detecting changes in workload

Detecting changes by measuring workload in a real-time setting was the final goal for this research. IWS and HRV measurements took place in a real-time setting and were gathered during the measurement sessions. XTL relies on ProRail logging data, this data was not available real-time, it took a few days before this data was available. This was no problem in this stage since is was not necessary to process the XTL data real-time.

A new metric called 'stretch' was introduced to compare XTL and IWS scores during a (cluster) of events. Stretches are basically changes in workload since they are marked by a rising of IWS and XTL scores. The stretch-ratio has been introduced as well, defined by the subjective- divided by the objective stretch and can be used to compare stretches to each other. A stretch with a higher stretch-ratio compared to one with a lower stretch-ratio indicates a higher subjective workload for the same objective workload and therefore a change in the relative workload.

10.4 Measuring workload of social-technical system in real-time

The main research goal was measuring workload of a socio-technical system in real-time setting. ProRail control post Zwolle was used as a proof of concept. Three measurements have been introduced to measure three types of workload; XTL, IWS and HRV. All of these measurements have been used in a real-time setting in an operational environment. IWS and HRV are not specific to ProRail and could be used in other control room settings as well. The IWS scale was developed specifically for rail signalers. The terms on the scale are tuned for rail signalers, this should be taken into account when it is used in some other field. XTL is very specific to ProRail since the input for this method is logging data from ProRail and the activities are specific for rail signalers. The concept of XTL can be used in other fields though, this would require to implement the components based on characteristics of that particular field.

11 Future work

This research is a first step for real-time measuring workload in an operational environment. The method developed can be extended and improved on a number of levels. This section provides some suggestions for this.

11.1 Changes IWS interface

The IWS interface has been used by all rail signalers during the two measurement sessions. The required input from rail signalers is limited, but they need to provide a score every 5 minutes. It is corrected when rail signalers forget to enter a score, in this situation the newly entered score is also filled in for the missing time slots. During the research we noticed that rail signalers were more willing to enter a score when their scores changed. So if a rail signaler has a base level of 2 and he does not enter a score for half an hour and he would enter a score of 3 the score of 3 would be entered for the missing interval as well. Since rail signalers are more willing to enter a score upon change it might be better to fill the missing scores with the last entered scores, so in this example fill in a score of 2 for the missing intervals. This requires that the rail signaler must fall back on the base-line by himself, otherwise the system will be filling the gaps with a higher score.

11.2 Determine XTL constants

The XTL constants for monitoring, planning, manual and communication actions are still set to 1. It makes sense to adjust these constants since the required effort for these actions is most likely different. From the XTL component distribution we can already see that communication plays a major part when the workload increases, so it therefore likely to increase the share of communication actions in the XTL formula by raising its constant. The relation between actions and effort need to be further investigated.

11.3 Real-time logging data

Logging data used to calculate the XTL score is extracted from ProRail logging data. The data is extracted from several sources and especially handling the phone logs cost some effort. An improvement will be when the this data becomes available real-time. This requires real-time access to ProRail logging. When this data becomes available in real-time it is also possible to calculate the XTL score real-time. This will result in real-time availability of the IWS and XTL scores.

References

[1] A. W. Siegel and J. M. Schraagen, "Developing resilience signals for the dutch railway system," *5th Resilience Engineering Symposium*, pp. 191–196, 2014.

[2] M. A. Neerincx, "Cognitive task load analysis: allocating tasks and designing support," in *Handbook of cognitive task design* (E. Hollnagel, ed.), vol. 2003, ch. 13, pp. 283–305, Mahwah, NJ: Lawrence Erlbaum Associates, 2003.

[3] J. Rasmussen, *Information processing and human-machine interaction. An approach to cognitive engineering.* North-Holland, 1987.

[4] D. D. Woods and J. Wreathall, "Stress-Strain Plots as a Basis for Assessing System Resilience," in *Resilience engineering perspectives, volume 1: Remaining sensitive to the possibility of failure* (E. Hollnagel, C. Nemeth, and S. Dekker, eds.), ch. 12, pp. 145–161, Aldershot, UK: Ashgate, 2006.

[5] L. Pickup, J. R. Wilson, S. Nichols, and S. Smith, "A conceptual framework of mental workload and the development of a self-supporting integrated workoad scale for railway signallers," in *Rail human factors* (J. Wilson, B. J. Norris, T. Clarke, and A. Mills, eds.), ch. 29, pp. 319–329, Ashgate, 2005.

[6] L. Pickup, J. R. Wilson, B. J. Norris, L. Mitchell, and G. Morrisroe, "The Integrated Workload Scale (IWS): a new self-report tool to assess railway signaller workload.," *Applied ergonomics*, vol. 36, pp. 681–693, Nov. 2005.

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

[8] S. Brennan, "An experimental report on rating scale descriptor sets for the instantaneous self assessment (isa) recorder'," *Portsmouth: DRA Maritime Command and Control Division. DRA Technical Memorandum (CAD5)*, vol. 92017, 1992.

[9] M. Wilms, M. E. Erg, and M. Zeilstra, "Subjective mental workload of dutch train dispatchers: Validation of iws in a practical setting," *4th International Conference on Rail Human Factor.*, pp. 641–650, 2013.

[10] Q. Gao, Y. Wang, F. Song, Z. Li, and X. Dong, "Mental workload measurement for emergency operating procedures in digital nuclear power plants.," *Ergonomics*, pp. 37–41, May 2013.

[11] G. E. Billman, "Heart rate variability - a historical perspective," *Frontiers in phisiology*, vol. 2, 2011.

[12] A. Hoover, A. Singh, S. Fishel-Brown, and E. Muth, "Real-time detection of workload changes using heart rate variability," *Biomedical Signal Processing and Control*, vol. 7, pp. 333–341, July 2012.

[13] A. D. Goedhart, S. van der Sluis, J. H. Houtveen, G. Willemsen, and E. J. C. de Geus, "Comparison of time and frequency domain measures of RSA in ambulatory recordings.," *Psychophysiology*, vol. 44, pp. 203–15, Mar. 2007.

[14] M. Malik, "Heart Rate Variability.," *Annals of Noninvasive Electrocardiology*, vol. 1, pp. 151–181, Apr. 1996.

[15] F. Togo and M. Takahashi, "Heart rate variability in occupational health –a systematic review.," *Industrial health*, vol. 47, pp. 589–602, Dec. 2009.

[16] L. Pickup, J. Wilson, and E. Lowe, "The Operational Demand Evaluation Checklist (ODEC) of workload for railway signalling.," *Applied ergonomics*, vol. 41, pp. 393–402, May 2010.

[17] J. A. Veltman and A. W. K. Gaillard, "Pilot workload evaluated with subjective and physiological measures," in *Aging and human factors* (W. C. M. J. Brookhuis, K and D. Waard de, eds.), pp. 107–128, University of Groningen, Haren, 1996.