Measuring Rail Signaller Workload in a Highly Realistic Simulated Environment

Birte THOMAS-FRIEDRICH\textsuperscript{1}, Philipp SCHNEIDER\textsuperscript{2}, Heiko HERHOLZ\textsuperscript{2}, Jan GRIPENKOVEN\textsuperscript{1}

\textsuperscript{1}Institute of Transportation Systems, German Aerospace Center (DLR), Germany
\textsuperscript{2}Department of Land and Sea Transport Systems, Technische Universität Berlin, Germany

Abstract. The perceived mental workload of rail signallers was assessed during three experimental scenarios in a simulated work environment using NASA TLX and ASWAT. Results indicate that a translation of ASWAT can be used to measure the mental workload of signallers in the German railway system. In the future, this will allow the comparison of signaller workload cross-culturally, enabling the comparison of e.g. different operational systems and their impact on workload.

Keywords. Mental workload, rail signalling, simulation, measurement

1. Introduction

The task of rail signalling is of crucial importance for the safety and punctuality of trains. Rail signallers operate points and signals and are responsible for the safe handling of disruptions. However, their job has been subject to a constant evolution. Enabled by new technologies and accelerated by companies’ cost pressure, routine tasks have been transferred to assistant systems like the automatic train control. This might have a major influence on signallers’ mental workload (MWL).

The job of rail signallers in Germany was born with the invention of mechanical interlockings, which were the first signal boxes in today’s sense. Operating a mechanical interlocking, signallers had the task of changing points and setting signals manually. This required a lot of physical labour and diligence, as many safety measures were implemented through visual checking by the signaler. An example is the necessity to check the tracks for other vehicles by eyesight before accepting a train movement. The development of electro-mechanical interlockings, relay and eventually electronic interlockings reduced the signallers’ physical strain, e.g. through electrical point machines and simplified setting procedures. A track clear detection system was developed, which provides complete information about the occupation and clearance of a track section, rendering the visual checking obsolete. Now, the amount of points and signals a signaller can control is not limited by the area s/he can monitor through a glance out of the window, instead, there are virtually no technical constraints how large a signaller’s area of responsibility can be (Bormet et al., 2017). Infrastructure managers have to define how large the area of responsibility can be that a signaller can safely handle.
The need for the dimensioning of signallers’ areas of responsibility is one reason why a reliable measurement of the MWL of signallers is necessary.

Further, the increase of automation in electronic interlockings influences signaller workload substantially. Automatic route setting systems were introduced. This automatic train control carries out large parts of rail signallers’ tasks during regular operations, e.g. the setting of signals or operation of points (Huke, 2017). The rail signaller has the mainly passive task of monitoring the automatic train control. Thus, rail signallers’ work has changed from mainly active work to mainly passive supervisory control. These monitoring activities are challenging for operators, especially with regard to keeping good situation awareness (Parasuraman, 1987; Endsley, 1996). The signaller has to monitor all events and to intervene abruptly in case of disruption or danger. During disruptions (e.g. a points failure), rail signallers thus face sudden intermediate periods of stressful and time critical active work. The rail signaller has to handle the disruption while at the same time ensuring safe rail operations with as little delays as possible (Thomas & Gripenkoven, 2016). Sudden changes between periods of passive, low-activity monitoring and periods of active, stressful work are likely to increase even more with a higher grade of automation, which infrastructure managers are eager to achieve (Guss, 2016). Hence, the MWL of rail signallers may vary substantially over a working shift, from a potential underload during normal operation to massive stress and overload with several disruptions simultaneously.

Both under- and overload and the sudden change of workload due to disruption can have detrimental consequences for the performance of operators (Young, 2015; Dobson, 2015). A non-optimal MWL can lead to human errors, attentional lapses and an overall degradation of performance. Optimal human performance is more likely to be achieved within an optimal range of MWL (Young, 2015). However, when analysing the effects of over- or underload with regard to workload, not only its influence on performance should be taken into account, but also the influence on the operators’ individual wellbeing (Pickup, Wilson, Sharple, Norris, Clarke & Young, 2005). In conclusion, an analysis of the MWL of rail signallers is necessary to enable the realistic dimensioning of areas of responsibility. Secondly, the analysis of MWL is necessary to gather information on how to achieve an optimal level of MWL, thus ensuring an optimal performance of rail signallers as well as mitigating negative impacts on operator wellbeing. Ensuring optimal human performance is highly relevant due to the importance of rail signallers’ work for the safety of the rail network. The present study aimed to facilitate the assessment of signaller MWL in German railway operations by determining if the use of the translated Adapted Subjective Workload Assessment Technique (ASWAT, Lowe & Pickup, 2008) is appropriate for application in this context.

1.1. Assessment of mental workload

As MWL is a concept with various definitions (Young, 2015; Pickup, Wilson & Lowe, 2010), it is necessary to determine which approach will be followed for the current study. Here, MWL was defined as perceived MWL, thus a subjective rating of the MWL by the operator was obtained in the current study. Perceived MWL was analysed because ensuring an optimal performance of the operator can be supported by a user-friendly work environment. To achieve such a user-friendly workplace, it is important to not solely focus on technology but to also understand the user or operator of a socio-technical system (Naumann, Gripenkoven, Giessenmann, Stein & Dietsch, 2013). There are numerous self-report tools used to assess perceived MWL of operators. In previous research regarding the perceived MWL of signallers, the NASA Task Load Index (NASA TLX; Hart and Staveland 1988) was used often. However, the use of the NASA TLX alone is not sufficient to assess signaller MWL, as it will not account for the domain specific factors influencing signaller MWL (Pickup et al.,
Measures that are tailored to the specific work environment of rail signallers are necessary. The current study focused on the ASWAT, a measure designed specifically for rail signallers (Lowe & Pickup, 2008). The ASWAT was developed as a rail-specific adaptation of the SWAT (Subjective Workload Assessment Technique; Reid and Nygren, 1988). The self-report measure assesses three dimensions of perceived MWL, time load, mental effort and pressure or stress in the context of rail signalling. The tool was developed to compare the perceived MWL between different situations retrospectively (Lowe & Pickup, 2008). The present study aimed to determine whether this measure can be used to assess MWL in the context of rail signalling in German railway operations.

2. Methods

In order to assess whether the use of the ASWAT is appropriate in the context of rail signalling in Germany, the measure was translated and applied during a simulator study. In addition, participants completed the NASA TLX as an established measure of perceived MWL. The following analysis was conducted: First, convergent validity of the translated ASWAT was analysed by correlating the results of ASWAT with results of NASA TLX. Second, objective workload was analysed using behaviour coding. Results of the behaviour coding were used to further examine the validity of the results obtained using ASWAT and NASA TLX. Third, the performance of signallers and the relationship between perceived MWL and performance was assessed during the experimental scenarios to gain a better understanding of the effects of changes in perceived MWL on signalr performance.

During the study, participants had the task of operating trains in an electronic interlocking. The interlocking was a new work environment for all participants, as the modelled infrastructure did not represent a real station. Before participating in the study, participants were sent a description of the experimental interlocking as well as the timetable used, to familiarize themselves with the workplace they would be asked to control. Upon arrival, participants received a brief explanation of the experimental interlocking, followed by a practice period. During this period, participants were instructed to handle all train traffic manually. They were able to ask questions and request further practice time.

Once the participants were familiar with the experimental environment, they completed three different experimental scenarios in a randomised order. In one scenario, participants had to operate the interlocking manually, directing all train traffic by changing points and operating the signals. In a second scenario, the automatic route setting system was active. Thus, signallers had the task of monitoring the smooth functioning of the system. In a third scenario, a sudden points failure occurred while the automatic route setting system was active. The signalr had to detect the disruption and act accordingly. The within-subjects experimental design allowed a comparison between the workload during passive work, active work and during the handling of a disruption. In the following, the scenarios are referred to as Manual, Automation and Disruption, respectively. Each scenario lasted for approximately 15 minutes. During the practice period and each scenario, the timetable remained the same. After each scenario, ASWAT and NASA TLX were completed. Participant’s behaviour was coded concurrently, while the participant was completing the experimental scenario.

2.1. Participants

Thirteen signallers took part in the study (male = 11). They were on average 25.08 years old (SD = 4.68). Nine signallers were currently actively working as signallers. They had on average 3.93 years of work experience (SD = 4.54). The four signallers who were not currently active had on average 0.33 years (4 month) of work experience (SD = 0.14). The
last time they actively worked was on average 1.1 years ago (SD = 0.56). Although they had not been working actively over a long period of time, they were completing their vocational training as part of a dual study program at a university of applied sciences and thus had expert knowledge of the railway system.

2.2. Measurements

ASWAT: The self-report tool measures three dimensions of perceived MWL (time load, mental effort and pressure) with one item per dimension. Each item has three answer options. For each option, a behavioural anchor is provided. Signallers can rate their perceived MWL on each scale as high, medium or low. The results are scored as 3, 2 or 1 point, respectively, and a sum score (ranging between 3 and 9) is calculated. The higher the sum score, the higher the perceived MWL.

NASA TLX: The workload dimensions of mental demand, psychical demand, temporal demand, frustration, effort and performance are assessed. The NASA TLX lets participants rate their MWL in each dimension on a continuous scale from 0 to 100. For a total score, the results for each item are averaged. During the development of NASA TLX it was established that participants had to rate the importance of each workload dimension for their task in addition to rating their perceived workload. In accordance with the recommendation of multiple authors this additional step was not conducted in the current study (Nygren, 1991; Hart, 2006).

Behavior Coding: For the behaviour coding the categories of behaviour used in previous studies (Balfe, Sharples & Wilson, 2015) were adapted. It was determined whether participants were monitoring the running of the trains, interacting with the work environment, dealing with the timetable or planning of the running of the trains, communicating via the telephone or having quiet time. Monitoring was differentiated in active and passive monitoring. Passive monitoring was defined as the participant taking a comfortable position, e.g. reclining in their chair, and observing the monitors of the work environment. Active monitoring was defined as the participant taking a more active position, e.g. leaning toward the monitors while supervising the functioning of the systems.

Performance: Signallers performance was rated on two dimensions. First, punctuality is an important indicator for the performance of a signaller. Due to technical constraints, punctuality could not be monitored constantly. Instead, the final position of trains at the end of the experimental scenario was used for the assessment, thus, punctuality was operationalised as percent of trains that were on time at the end of the scenario. In addition to punctuality, an expert assessment of signaller performance was obtained. An expert rated each signaller right after completing the experimental scenario on a German school grade scale from 1 to 6, 1 being the best and 6 equalling a failing grade.

2.3. Experimental environment

In Germany railway operations laboratories – real interlockings with model railways as output medium – have a long tradition and play an important role in education. The Railway Operations Laboratory (Eisenbahn-Betriebs- und Experimentierfeld, EBuEf) at Technische Universität Berlin is used in trainings for engineers, students and signallers. The didactic benefit is described by Friedrich et al. (2016). Students can use several mechanical, electromechanical and relay interlockings, operating model trains on a network length of some 29km (Blome, 2011). One unique feature is the electronic interlocking “WebSTW”, the functionality and appearance of which are very close to its industry counterparts in service for DB Netz. However, as it is an in-house development, it can easily be adjusted to new requirements, e.g. new infrastructure or modified appearances. This makes it an ideal platform
for human factors experiments. The WebSTW workplaces within the EBuEf are separated from the remaining low-tech signal boxes and the model railway. Nevertheless, a real model train moves when the signaller sets a signal via WebSTW. As in a real operations centre, the signaller can visually follow the train movement only on his screens and communicate with other entities only by phone. For the current study, the EBuEf infrastructure was adapted to provide one workplace with an electronic interlocking displayed on six screens, via which 10 stations, 35 points and 44 signals were to be operated on a predominantly single-track network. In addition, timetable information was supplied on a seventh screen. The work environment is depicted in figure 1.

![Figure 1: Work environment](image)

### 3. Results

#### 3.1. Workload

The results for the assessment of MWL using the NASA TLX and ASWAT are shown in Figure 2. For both measures the total score was used for the assessment of perceived MWL.

![Figure 2: Boxplots of the perceived mental workload assessed with NASA TLX and ASWAT](image)

Repeated-measures ANOVAs were conducted to analyse whether the workload differed significantly between scenarios. With regard to the results of the ASWAT, the scenario had a significant effect on the perceived workload \(F(2, 24) = 4.85, p = .017\), with a medium effect size \(\eta_p^2 = 0.18\). A significant effect of the scenario was also observed with regard to the results of NASA TLX \(F(2, 24) = 3.56, p = .044\). Here, the effect size was small to medium \(\eta_p^2 = 0.12\).
Post hoc paired comparisons using the Bonferroni-Holm correction (Holm, 1979) showed that for the results of the ASWAT, there was a significant increase of MWL during the disruption compared to the automated scenario ($M_{\text{auto}} = 4.39$, $SD_{\text{auto}} = 1.19$; $M_{\text{disr}} = 5.85$, $SD_{\text{disr}} = 1.28$; $p = .044$). All other differences were not significant (see Table 1 for M and SD).

Post hoc paired comparisons for the results of NASA TLX showed no significant differences between the scenarios, although the MWL between the automated scenario and the disruption increased by 14 points ($M_{\text{auto}} = 32.67$, $SD_{\text{auto}} = 17.98$; $M_{\text{disr}} = 46.33$, $SD_{\text{disr}} = 14.92$; $p = .06$). The non-significance may in part be due to the small sample size.

### Table 1: Mean and Standard deviation of the workload and performance for each scenario

<table>
<thead>
<tr>
<th>Condition</th>
<th>M</th>
<th>SD</th>
<th>Condition</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>NASA TLX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual</td>
<td>39.21</td>
<td>14.68</td>
<td>Manual</td>
<td>78.21</td>
<td>19.70</td>
</tr>
<tr>
<td>Automation</td>
<td>32.67</td>
<td>17.98</td>
<td>Automation</td>
<td>94.87</td>
<td>8.01</td>
</tr>
<tr>
<td>Disruption</td>
<td>46.33</td>
<td>14.92</td>
<td>Disruption</td>
<td>71.79</td>
<td>12.52</td>
</tr>
<tr>
<td>ASWAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual</td>
<td>5.31</td>
<td>1.55</td>
<td>Manual</td>
<td>2.39</td>
<td>0.77</td>
</tr>
<tr>
<td>Automation</td>
<td>4.39</td>
<td>1.19</td>
<td>Automation</td>
<td>2.23</td>
<td>1.09</td>
</tr>
<tr>
<td>Disruption</td>
<td>5.85</td>
<td>1.28</td>
<td>Disruption</td>
<td>2.62</td>
<td>1.04</td>
</tr>
</tbody>
</table>

3.2. **Assessment of ASWAT**

Both measures were correlated using Spearman Rho-correlations to assess convergent validity. The results are shown in Table 2. Both measures were highly correlated and all correlations were significant, showing that both tools seem to be measuring the same construct, perceived MWL.

### Table 2: Correlations of NASA TLX and ASWAT for each scenario and all scenarios

<table>
<thead>
<tr>
<th></th>
<th>All scenarios</th>
<th>Manual</th>
<th>Automation</th>
<th>Disruption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>.8**</td>
<td>.78*</td>
<td>.77*</td>
<td>.84**</td>
</tr>
<tr>
<td></td>
<td>* = p &lt; .05;</td>
<td>** = p &lt;.001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 3: Correlations of the items of the NASA TLX and the ASWAT

<table>
<thead>
<tr>
<th>Item</th>
<th>TLX&lt;sub&gt;Mental&lt;/sub&gt;</th>
<th>TLX&lt;sub&gt;Physical&lt;/sub&gt;</th>
<th>TLX&lt;sub&gt;Temporal&lt;/sub&gt;</th>
<th>TLX&lt;sub&gt;Effort&lt;/sub&gt;</th>
<th>TLX&lt;sub&gt;Frustration&lt;/sub&gt;</th>
<th>TLX&lt;sub&gt;Performance&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASWAT time load</td>
<td>0.66**</td>
<td>0.25</td>
<td>0.77**</td>
<td>0.73**</td>
<td>0.32*</td>
<td>0.22</td>
</tr>
<tr>
<td>ASWAT mental effort</td>
<td>0.63**</td>
<td>0.36*</td>
<td>0.55**</td>
<td>0.57**</td>
<td>0.56**</td>
<td>0.57**</td>
</tr>
<tr>
<td>ASWAT pressure</td>
<td>0.19</td>
<td>0.16</td>
<td>0.32*</td>
<td>0.43*</td>
<td>0.61**</td>
<td>0.65**</td>
</tr>
</tbody>
</table>

* = p < .05; ** = p >.001

Items in both measures assessing the same dimension of perceived MWL were expected to correlate. Correlations of the single items of both measures were conducted to analyse this
relationship. All correlations are shown in Table 3. As expected, ASWAT \textsubscript{Time load} correlated most strongly with TLX \textsubscript{Temporal Demand}, but also with TLX \textsubscript{Effort} as well as TLX \textsubscript{Mental Demand}. ASWAT \textsubscript{Mental effort} correlated with all items of the NASA TLX except for TLX \textsubscript{Physical demand} in a fairly equal range, with the largest correlation being with TLX \textsubscript{Mental demand}. ASWAT \textsubscript{Pressure} correlated most strongly with TLX \textsubscript{Performance} and TLX \textsubscript{Frustration}. Overall, the correlations that were to be expected (e.g. mental demand and mental effort) were found, further supporting that similar dimensions of perceived MWL were assessed by both measures.

3.3. Behaviour coding

The results for the behaviour coding are shown in Figure 3. Passive monitoring (i.e. the participant observing the monitors in a comfortable position) was displayed more frequently with the introduction of the automatic route setting system. It was expected that participants would display passive monitoring most frequently during the automated scenario; however, passive monitoring was displayed most often during the disrupted scenario. This might be due to the fact that the disruption was only introduced during the last part of the scenario. Participants were displaying all coded behaviours more frequently during the disrupted scenario in comparison to the automated scenario, resulting in a higher objective workload. Thus, the results of a higher perceived MWL during the disrupted scenario compared to the automation are supported. The MWL in the manual scenario, however, was not perceived to be higher than that in the other two scenarios, even though active monitoring, planning and interaction with the system were displayed most frequently during the manual scenario. All three behaviours were displayed least often during the automated scenario, as was expected. It seems that the more active task during the manual scenario was not perceived as significantly more challenging than the monitoring task during the automated scenario.

3.4. Performance

The results for the assessment of the performance, i.e. expert assessment and punctuality, are shown in Figure 4.
Similarly to the analysis of the workload, repeated-measures ANOVAs were conducted. The punctuality differed significantly between the scenarios \( F(2, 24) = 9.17, p = .001 \), with a large effect size \( (\eta^2_p = .34) \). Post hoc paired comparisons using the Bonferroni-Holm correction showed that the punctuality in the automated scenario was significantly better than in the manual and disrupted scenario \( (p = .041; p = .000, \text{ respectively}; \) for the M and SD see Table 1). The punctuality between the manual and disrupted scenario did not differ significantly. Overall, the results have to be interpreted carefully; firstly, because a certain amount of delays was unavoidable during the disrupted condition. Secondly, a superior punctuality due to automatic route setting was expected.

With regard to the expert assessment, no significant difference between the scenarios was observed \( F(2, 24) = 0.85, p = .438 \). This indicates that the order in which the participants took part in the experiments did not influence their performance, as they were able to perform equally well over all three scenarios.

3.5. Correlation of workload and performance

Spearman-Rho correlations were conducted. Punctuality did correlate negatively with the results of the ASWAT and NASA TLX, both correlations were significant. The correlations with the expert assessment were not significant (see Table 4).

<table>
<thead>
<tr>
<th></th>
<th>Punctuality</th>
<th>Expert assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASWAT</td>
<td>-0.48*</td>
<td>0.23</td>
</tr>
<tr>
<td>NASA TLX</td>
<td>-0.37*</td>
<td>0.09</td>
</tr>
</tbody>
</table>

4. Discussion and conclusion

The overall goal of the current study was to analyse whether a translation of the ASWAT can be used to effectively assess perceived MWL of rail signallers in Germany. Results show that both measures correlate strongly and significantly, both with regard to the total scores but also on item level. This indicates convergent validity, in that both measures assess perceived MWL. Thus, it can be assumed that the translation of the ASWAT can successfully be implemented for use in the German railway system.

Results indicate that both measures have certain advantages with regard to the question which measure should be preferred for the application in the context of rail signalling in Germany.
Workload differed significantly between the automated and disrupted scenario when measured with the ASWAT. Behaviour coding showed similar differences between those scenarios as well. Pairwise comparisons for the results of the NASA TLX did not show significant differences. Thus, the ASWAT is able to differentiate the workload better between the scenarios than the NASA TLX. However, the non-significant differences in the results of the NASA TLX might also be due to the small sample size. The NASA TLX has the advantage of assessing more dimensions of perceived MWL separately, as multiple items of the NASA TLX correlated highly with each of the items of the ASWAT. Thus, the NASA TLX allows for a more nuanced analysis of perceived MWL. However, the NASA TLX takes longer to complete and is not easy to understand. In practice, depending on the purpose, researchers should decide which measure would be best suitable to provide the desired information.

Interesting results could be observed with regard to the behaviour coding. Even though planning, interaction with the system and active monitoring were displayed most frequently during the manual scenario, the perceived MWL in the automated scenario did not differ significantly from that in the manual scenario. These results might be due to the fact that monitoring consistently is also a highly demanding task (Young, 2015). As the load due to active tasks lowered, the perceived MWL might have stayed the same as signallers had to monitor the functioning of the system and understand the automations actions. The observed results differ from those obtained by Balfe, Sharples and Wilson (2015). In a study with six signallers, they showed a decrease in perceived MWL when automation increased. This might be due to the differing way the automation was implemented in both studies and could point to interesting effects of different types of automation on perceived MWL. This should be investigated further in future research with a larger sample size and a direct comparison. The results with regard to signaller performance should be interpreted carefully, as e.g. the punctuality was not assessed over time but only at the end of each scenario. In addition, punctuality was influenced by the type of scenario, with the disruption leading to unavoidable delays and the automation making delays very unlikely. In accordance with Balfe, Sharples and Wilson (2015), more detailed and varied measures of performance would be necessary to allow for a thorough interpretation of performance effects. Still, differences in perceived workload correlated significantly with the punctuality. It is possible that signallers noticed the amount of delays they were producing and their assessment of their workload was thereby influenced. Overall, the relationship between perceived MWL and performance should be analysed further with a more in depth analysis of performance.

Limitations of the study are due to the fact that data collection in real-life interlockings is difficult to achieve. Thus, data was collected in a simulated work environment which signallers did not know beforehand. Consequently, signallers in the study were lacking the expertise a signaller has for “his/her” interlocking after working with it over a period of multiple years.

In conclusion, this study demonstrates a first approach to validate the use of the ASWAT in the context of rail signalling in German railway operations. It was shown that the ASWAT can be applied successfully in this context. In the future, this will allow the comparison of results regarding rail signaller workload cross-culturally. Thus, it would be possible to compare e.g. the impact of automation on rail signaller workload across different operational systems and concepts of automation, especially with regard to the question of over- and underload and its consequences for signaller performance. The results could be used to improve rail signallers’ work environments in terms of their user friendliness, adaptive automation and interface design.
Acknowledgements: The authors would like to thank Anni Drescher and Anton Lichtenberg for their valuable support during the realization of the current study.

References


