

2017

How Useful are Carboard Mock-Ups: the Use of Different Levels of Simulation Fidelity in Assessing Signallers' Workload

Andrew Marshall

Marshall Ergonomics Ltd., andy@mel.bz

James Murphy

HAM Associates Ltd., jamesm@ham.co.uk

Follow this and additional works at: <https://arrow.tudublin.ie/hwork17>



Part of the [Computer Sciences Commons](#)

Recommended Citation

Marshall, A. & Murphy, J. (2017). How useful are cardboard mock-ups?: the use of different levels of simulation fidelity in assessing signallers' workload. *H-Workload 2017: The first international symposium on human mental workload*, Dublin Institute of Technology, Dublin, Ireland, June 28-30. doi:10.21427/D7SS55 isbn:9781900454637

This Conference Paper is brought to you for free and open access by the H-Workload:Models and Applications at ARROW@TU Dublin. It has been accepted for inclusion in H-Workload 2017: Models and Applications (Works in Progress) by an authorized administrator of ARROW@TU Dublin. For more information, please contact yvonne.desmond@tudublin.ie, arrow.admin@tudublin.ie, brian.widdis@tudublin.ie.



This work is licensed under a [Creative Commons Attribution-Noncommercial-Share Alike 3.0 License](#)

How Useful Are Cardboard Mock-Ups? The Use of Different Levels of Simulation Fidelity in Assessing Signallers' Workload

Andrew Marshall¹ and James Murphy²

¹ Marshall Ergonomics Ltd, 25 Third Avenue, Havant, PO9 2QR, UK andy@mel.bz
² HAM Associates Ltd, Black Prince Road, London SE1 7SJ, UK jamesm@ham.co.uk

Abstract. Two techniques were utilised: the Bedford Scale and the NASA TLX. Assessments were made with two levels of fidelity. The first used a busy 15 minute scenario with seven different failure conditions using paper based layouts of the new signalling system. The second used a three dimensional representation of the proposed signal box layout for a busy one hour scenario. A key finding was that the new box could be run by two signallers with acceptable levels of workload, even with minor failures. A number of changes to the layout were proposed based on the experience of an hour's simulation. The methodology showed that a cardboard model can be a useful tool in a participatory approach aiding the assessment of design and mental workload in a format that can be readily understood by all the stakeholders.

Keywords: Mental Workload, Simulation, Rail Signalling

1 Introduction

The workload assessment studies referred to here were carried out in 2002 as part of a larger project to rationalise three Lincoln signal boxes (SBs), with the option of a fourth, into one SB.

The background to the project is given, together with explanations of the methodologies adopted. The available techniques for assessing signaller workload have advanced markedly since the project and these will be discussed with a view to what the authors might select for a present day study. It will be argued that the use of low fidelity simulations, to facilitate examination of the design and workload of an operational environment, are still applicable today.

2 Background

This project was part of a scheme to rationalise the Lincoln area into a single control area, controlled from Lincoln High Street SB. Railtrack had also requested that the control panel should be easily extendable in the future to incorporate the West

Holmes control area. The workload study was undertaken as part of the project to predict if it could be run effectively by one or two signallers.

All the SBs were all equipped with lever frames and block shelves with the signallers accepting and passing on trains via the use of block instruments. Lever frames are mechanically connected by wires to points and signals and are manually operated. Fig. 1 is a photo of Lincoln High Street SB.



Fig. 1. The lever frame at High Street

The frames, shelves and block instruments were to be replaced with an NX (eNtry eXit) signalling control and indication panel with individual in-line TD (Train Descriptor) stepping berths. NX panels have a schematic representation of the track and signals with trains being routed using entry and exit buttons on the panel. Setting the start (entry) and end (exit) defines a route for a train and sets all the intermediate interlocking (points and signals) accordingly. Fig. 2 is an example of an NX panel.

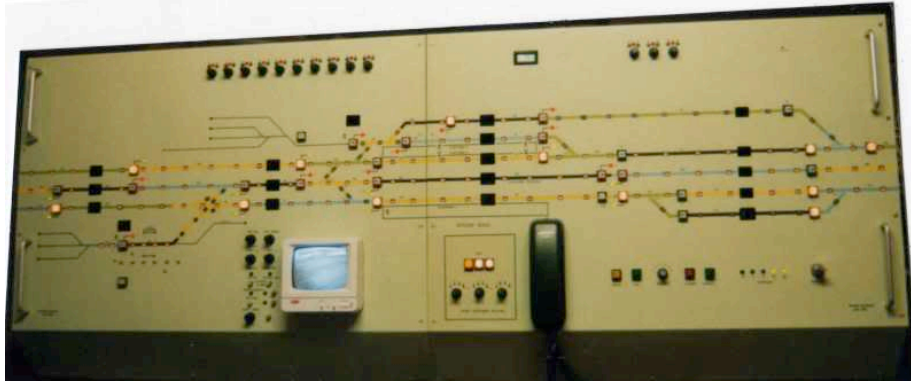


Fig. 2. An example of an NX panel

A task analysis was carried out for working under absolute block regulations with bell communications (the then current method of working in these SBs). Another task analysis was performed for Track Circuit Block operations using NX Panels. The signaller's main goal is to progress trains efficiently and safely. Millen and her colleagues [1] produced an abstraction hierarchy for railway signalling that is a more generic analysis of a signaller's tasks.

The task analyses were used to produce a SHERPA analysis (Systematic Human Error Reduction and Prediction Approach, see Embrey [2]) to identify potential errors and remedies. The predicted errors for current operations were more numerous than for the NX panel working. Errors under Track Circuit Block at worst tended to lead to delays, as the system fails safe. The exception is if level crossings are not clear and this is not perceived.

This paper concentrates on the workload analyses; the other activities are outlined to give some context of the user centred approach that was adopted.

3 Workload Assessment

3.1 Assessment Methods

A large variety of workload assessment techniques existed at the time. However, as Pickup et al wrote in 2005 [3] "To say that mental workload has been studied so extensively in other industries including transport (e.g. in aviation) there is a dearth of contributions to the human factors literature on workload and the railways".

Workload measures can be grouped into five categories, as described by Meshkati, Hancock and Rahini [4]:

- Primary task measures
- Secondary task measures
- Physiological measures
- Operator Modelling
- Subjective measures

3.2 Primary Task Measures

These are measures of job performance or effectiveness. In this context they could include items such as errors and delays to the timetable. These measures relate solely to the output of the operator and do not show the internal “cost” to maintain such performance. Primary task measures are often used with secondary measures to try and establish what margin of resource is still available.

Primary task measures require a working system or a high fidelity simulation. They also require validation to show sensitivity to task loads. Variations from the road centre line or a flight path are good examples of primary task measures. Control movements, and more particularly reversed control movements, also offer a source of data that can reflect the skill or attention being given to the task. Errors are usually infrequent and so may not be captured at all or be too few to reliably indicate load.

Primary task measures were not suitable here as we did not have a simulator and had no validated measures. Also they may not, of themselves, offer information about the load on the operator.

3.3 Secondary Task Measures

These are extra tasks given to the operator to try and determine the amount of spare mental capacity that is available while he is performing his usual tasks. The operator is instructed to concentrate on the primary task, but do the secondary task as well as he is able. Cancelling a light that comes on at random is an example. The delay in cancelling the light and any missed lights are the dependent variables that are measured to indicate spare capacity. A secondary task that showed high ecological validity in a number of scenarios was a prospective memory task, Sauer [5]. For example, the operator has to read a figure (e.g. a fuel gauge) at fixed intervals, such as 3 minutes. Deviations from the fixed time, missed readings and errors are again the dependent variables.

Secondary task measures do give information on the load on the operator, and together with the primary measures, the performance of the system. However, the drawbacks of the primary measures, such as resources required, still hold.

3.4 Physiological Measures

Physiological measures to indicate workload come in a broad range of techniques. All the physiological measures depend on doing the actual task or using a high fidelity simulator. Some of the measures require expensive equipment, time to get baseline levels and analyse results. Some of the measures are quite invasive – e.g. blood samples, temperature probes. They are, in general, not very good at differentiating workload levels on their own.

3.5 Operator Modelling

Models of the human operator at the time included:

- VACP (Visual, Auditory, Cognitive and Psychomotor loading, McCracken and Aldridge [6])
- IPME (Integrated Performance Modeling Environment, Farmer et al [7])
- CREWCUT (Dahl, et al [8])

One advantage of these models is that they enable predictions of workload for systems that have not yet been built to be examined. All are built on a detailed task analysis of the system under consideration. They are mostly very time consuming so were not suitable in this context.

3.6 Subjective measures

There are a plethora of subjective workload measures. Many have their history in the aviation industry. The list below includes some of the more common measures:

- Cooper Harper
- Systems Technology Scale
- Bedford Scale
- NASA TLX (Task Load Index)
- SWAT (Subjective Workload Assessment Technique)
- DRAWS (DERA Workload Scale)

The Cooper Harper scale was initially produced to help standardise the assessment of aircraft handling characteristics [9]. There are many similar scales to the Cooper Harper, such as the Systems Technology Scale. Both the Cooper Harper and the System Technology scale examine workload in relation to how much operator effort is required to overcome inherent problems in the design of the system. The Bedford Workload Scale is designed to assess an operator's spare mental capacity.

Uni-dimensional scales do not examine the sub components of workload, and therefore have limited diagnostic power for helping to produce improvements if a system is thought to have a workload problem. Multi-dimensional workload scales do have diagnostic capabilities. These scales include NASA TLX, SWAT and DRAWS.

SWAT, the Subjective Workload Assessment Technique (Reid and Nygren, [10]) is a subjective rating technique that uses three levels: (1) low, (2) medium and (3) high, for each of three dimensions of time load, mental effort load, and psychological stress load to assess workload. Its great advantage is that a trained operator can give a SWAT rating very quickly while carrying out his tasks. These can be verbal in response to a (verbal) prompt or at a fixed time. He only has to say for example "two, two, one".

SWAT requires three procedures for each participant and set of tasks or job. The first is scale development. All possible combinations of three levels of each of the three dimensions are contained in 27 cards. Each operator sorts the cards into the rank order that reflects his perception of increasing workload. This takes about 45 minutes. Conjoint scaling procedures are used to develop a scale with interval properties. The second procedure is the actual rating of workload for a given task or mission segment.

In the final step, each three-dimension rating is converted into numeric scores between 0 and 100 using the interval scale developed in the first step.

The NASA TLX (Hart and Staveland [11]) has six dimensions to assess mental workload: mental demand, physical demand, temporal demand, performance, effort and frustration. A weighting procedure is used to combine the six individual scales.

There is some evidence that the TLX and SWAT procedures can be simplified by removing the weighting from TLX and the sorting from SWAT. Byers, Bittner and Hull [12] compared SWAT with and without using the weighting procedure. Rubio, Martín and Díaz [13] compared the benefits of two different ways to computing the overall workload scores for a particular task when using TLX or SWAT. Two global scores were obtained for each task when using TLX: weighted (WTLX) (weighting the rating scores as above) and unweighted TLX (using a mean of the subscale ratings). Also, two SWAT overall scores were calculated for each task: using the same average scale for all subjects (GSWAT) and using the scale obtained for each group (SWAT) depending on the dimension which was the most important for the subject.

The results showed high correlations between the two TLX scores in all task conditions, but not for SWAT. However, the two TLX scores and the two SWAT scores showed similar sensitivity to task difficulty. In 2002 there was not enough data to warrant leaving out the time consuming procedures from these two techniques.

DRAWS has four dimensions of measurement. The rating scales are input demand (demand from the acquisition of information from external sources), central demand (demand from mental operations), output demand (demand from the responses required by the task), and time pressure (demand from the rate at which tasks must be performed). DRAWS ratings are easy to obtain, like SWAT although the scales are from 0 to 100. Ratings of more than 100 are also allowable if the demand is higher than can be coped with.

Subjective measures are easy to obtain. They can be inaccurate as subjects can perceive that they are working harder or less hard than other measures would indicate. They are prone to differences between subjects (inter-rater variability) and differences between the same subjects on different days (intra-rater variability). However they have high face validity, and can be used to predict workload for a system that does not yet exist.

3.7 Selected Methods

The Bedford Scale was selected for the first stage of the project. It is simple to administer easily understood. It is a proven scale having being used for over 20 years (a review of its effectiveness after ten years was given by Rosco and Ellis [14]). If workloads are within acceptable boundaries then the diagnostic power of multi-dimensional tools are not required. The Bedford Scale was adapted slightly for this project by replacing the word pilot with operator for ratings of 10. The scale is shown in Figure 3. If workload is predicted to be too high from the Bedford Scale (7 to 10) then some diagnostics might be required. This was planned to be either the NASA TLX or DRAWS. SWAT was felt to be over demanding in the time required to produce an interval scale for this project.

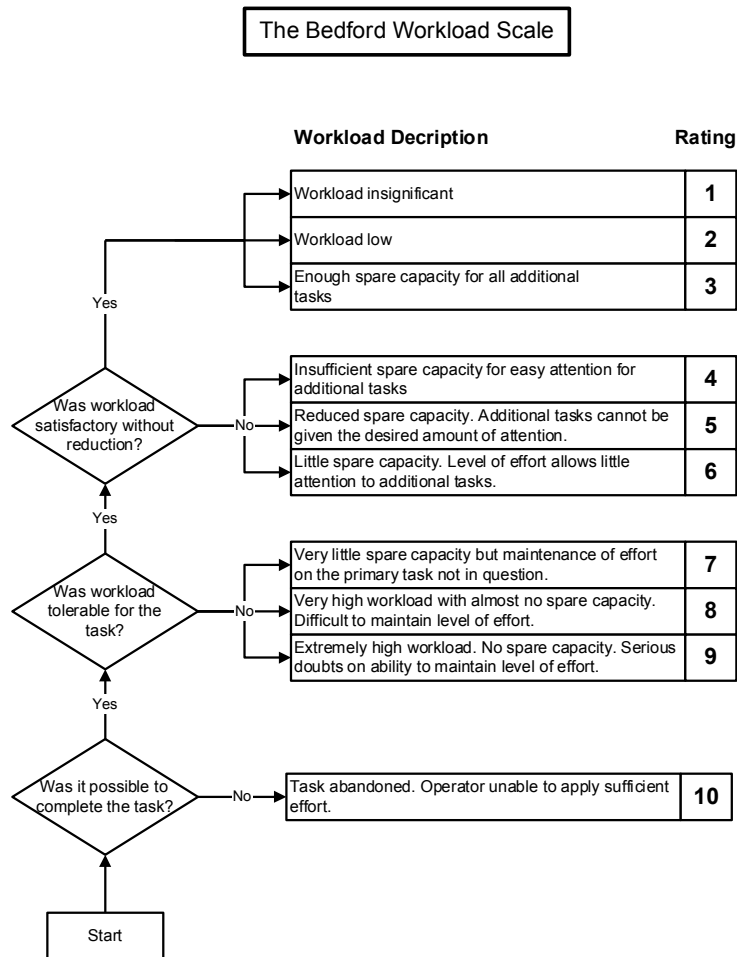


Fig. 3. The Bedford Workload Scale

4 Study 1

4.1 Scope

The first study included a task analysis for a lever frame under absolute block operation (the current signalling system) and an NX panel under track circuit block, a human error analysis and a desktop/paper based workload study. Only the workload study is reported here.

4.2 Workload Study and Findings

Workload was predicted using the modified Bedford Workload Scale. Train movements for a particular day were used to identify a busy period (see Table 1). It has a variety of movements including a train that terminates and then has to be moved to another platform (with the current track layout, but not for the new signalling scheme) to form a different service, leaving in the same time period.

Table 1. Train Movements between 10:15 and 10:29

| Arrive | Depart | Identify | Origin | Depart | Destination | Arrive | Type | Days |
|--------|--------|----------|-----------|--------|-------------|--------|------|------|
| 10:15 | ***** | 2P59 | HUDDFIELD | 07:07 | LINCLNCEN | 10:15 | WTT | SX |
| ***** | 10:20 | 2K15 | LINCLNCEN | 10:20 | PETERBORO | 11:52 | WTT | SX |
| 10:20 | 10:22 | 2J39 | GRIMSBYTN | 09:30 | SHREWSBRY | 14:02 | WTT | SX |
| PASS | 10:22 | 4K68 | WBURTONPS | 09:43 | IMM NCBP1 | 11:18 | STP | SX |
| 10:25 | 10:27 | 2E56 | SHREWSBRY | 06:52 | GRIMSBYTN | 11:24 | WTT | SX |
| ***** | 10:27 | 2B64 | LINCLNCEN | 10:27 | HUDDFIELD | 13:04 | WTT | EWD |

Failure modes were also identified and rated with the SMEs to assess the workload. The worst case situations to be chosen were track circuit, point and signal failures.

Two SMEs (Signal Managers) performed a “walk through” of the 15 minute scenario. Paper representations of the trains were moved on the new track diagram. Workload scores were then collected for operations with one and two signallers on different shifts and for failure conditions.

The exercise was repeated two weeks later with two Signalling Managers (one from the previous exercise and one other). However, this time the train movements, barrier operations and route setting were examined for each minute in the scenario. An updated track layout was also used. This brought to light that there was a conflict between two trains (2K15) and (2J39) at 10:20. In addition the Short Term Scheduled Freight train (4K68) would also be passing the same set of points in a conflicting direction at 10:22.

Following the second exercise, with the more detailed scenario, the SMEs were asked to confirm the ratings for each condition. These ratings are given in Table 2.

Table 2. Workload ratings for busy 'normal' period

| Condition | Combined 3 areas | Combined 4 areas |
|---|------------------|------------------|
| 2 Signallers - Scenario as in table 2 | 7/8 | 4 |
| 1 Signaller - timings of 3 trains that clash rearranged | 6/7 | |
| 2 Signallers - Night Shift | 3 | 3 |
| 1 Signaller - for 15 minutes | 9 | |
| 1 Signaller - for the day shifts | 9 | |
| 1 Signaller – for a night Shift | 7 | 6/7 |

At the first meeting the SMEs assumed that two signallers would divide the areas of responsibility in half. It was subsequently suggested that one would be responsible for the High Street Level Crossing (LX), with direct window sightlines from the SB and the other for the operation of the NX Panel. They could exchange duties as required. It was later affirmed by the SMEs that two signallers could also manage the larger control area (with the addition of West Holmes) if the High Street LX were to be converted to an MCB CCTV (Manually Controlled Barrier with Closed Circuit Television).

Lower workload ratings were given when it was assumed that the signalling for West Holmes is incorporated within the same SB. This is despite two extra CCTV crossings and the extra routing of trains that would be required. The SMEs believed that the extra notice time that they would have for each train, and the flexibility of being able to use the slow (currently goods) line to separate trains, would significantly reduce their workload. It would also reduce the communications required to co-ordinate the use of the slow line.

The predicted workload ratings for potential failures are very high. Although they are mostly 9/10 the SMEs said that the task would never be abandoned (WL 10). The high workload comes partly from the increase in communications load that would result from a failure, see Table 3.

Table 3 Workload ratings for single failures

| Condition | Workload Rating |
|--|-----------------|
| Barrier Failure (out of town) | 9/10 |
| Barrier Failure (city centre) | 9 |
| Train Describer Failure | 9/10 |
| Resetting axle counters | 6/8 |
| Track Circuit Failure | 9/10 |
| Point Failure 4846A | 9/10 |
| Train Failure (at points) or where Single Line working would have to be used | 9/10 |

The signallers in the High Street SB would be the focal point for co-ordinating the repair work for the failures. Putting extra staff in the Signal Box at this time would not reduce the load, as it is the signaller who has the overall picture of what is happening and needs to control it directly. This includes having to communicate with drivers, staff on the ground, and contractors that are repairing the fault. Most of these faults would result in delays to trains and then the routine of the timetable is disrupted for an extended period. The SMEs felt that the very high workloads would extend for the duration of the failure.

The city centre barrier failure is rated as slightly lower than an out of town failure. If the High Street barrier fails then the signaller in charge of the barrier could operate it manually although a long barrier failure would require assistance from other staff. Failure of other LXs would mean having to slow or stop trains to ascertain that the crossing is clear until the barrier can be staffed.

Failures in general, will not change how the Signal Box operates currently. One exception is track circuit failures that would cause reversion to Temporary Block

Working (similar to Absolute Block) for part of the signalling area. However, there will be two signallers to cope with the failure, rather than the one presently.

4.3 Study 1 Conclusions

Workload for a single operator was judged to be too high for sustained operations during the busy 15 minute period. The workload predictions given for operation by two signallers were at the boundary of acceptable levels under normal operating conditions (WL 6/7) for the highest workload period. Adding more signallers would not reduce this level. However, ratings for the night shift with two signallers were satisfactory without reduction (WL 3).

Methods of giving advanced warning of trains coming from the West Holmes area was recommended to provide more time for the signallers to manage their workload. If West Holmes were to be included into the current project timescales, this requirement would not be necessary.

Under single failure conditions the workload is predicted to be extremely high for an extended period of time and these might have to be maintained for several hours. This time period is usually dependent on response times of the external contractors involved. There also appears to be a shortage of staff available to cope with perturbed situations.

The predicted workload ratings will be validated during the simulations trials when a full scale model of the Signal Box will be constructed. Using a longer time frame (1 hour) would give more representative task loadings in conjunction with a multi-dimensional workload scale (NASA-TLX) to analyse various components of workload and help diagnose any problems.

5 Study 2

5.1 Introduction

It was recommended that the initial results required further evaluation during simulation trials with a full-scale mock-up of the Signal Box to provide a more reliable setting and employ a multi-dimensional workload rating scale to help diagnose sources of high workload.

A full-scale flexible mock-up of the Signal Box was constructed, which included the NX panel, face plate drawing and all the main equipment items, job aids and storage. A busy one hour scenario was developed to be used during the Simulation Trials. This was based on 19 train movements in the hour.

5.2 Methodology

Physical Arrangement. A full-scale flexible mock-up of the Signal Box was constructed which included the NX panel and all the main equipment items, job aids and storage. A full-size drawing of the NX panel's faceplate was used to show

control and display positions and allow train movements to be denoted. The mock-up was made of cardboard with a faceplate drawing printed out and adhered to the cardboard model. Train positions and TD codes were produced that could be stuck on at appropriate positions and moved.

The Simulation trials were based on the busy one hour scenario. The ergonomics study team moved the train position indicators (shown as red strips) along the tracks on the panel. The head codes for each train were put in the associated in-line berths (leading the progress of the train). The train positions and associated head codes were updated every minute. The signallers simulated setting routes, operating level crossings, and inputting Train Descriptions for trains originating from Lincoln. They also communicated with each other to pass information on train movements, barrier positions, etc.

Workload Measures. The NASA-TLX methodology was used to investigate workload and to help diagnose underlying causes of high workload predicted from the previous study. The Bedford Workload Scale was also used to help compare with the previous report and to obtain fast feedback on perceived workload.

The NASA-TLX has six dimensions to assess mental workload: mental demand, physical demand, temporal demand, performance, effort and frustration. A score from 0 to 100 is obtained on each scale. A weighting procedure is used to combine the six individual scale ratings into a global score. The operator is then required to make paired comparisons to select which dimension is more relevant to the workload between all pairs of the six dimensions. The number of times a dimension is chosen as more relevant is the weighting of that dimension scale. A total workload score (0-100) is produced by multiplying each scale rating by its weighting, summing across scales, and then dividing by 15 (the total number of paired comparisons).

Scenarios. A busy schedule of one hour's worth of train movements was devised for the study. There are 20 trains in the scenario, one disappears towards Newark after the first minute and one is at platform 4 at the start of the scenario and forms another service. There are therefore 18 active train movements in 60 minutes. The scenario is relatively quiet for the first quarter of an hour, progressively becoming busier, and easing off slightly in the last period.

Participants. Three Signalling Managers were used as participants in the study, two acted as signallers and one as an observer and advisor. One signaller controlled the West Holmes control area in the first trial and operated the High Street barrier with the LX pedestal in the second trial, and the second signaller controlled the remainder of the signalling area.

Procedure. The ergonomics study team (3 people) practiced the sequences for moving the trains and head codes for the one-hour scenario. They used timetables indicating train positions and highlighting the trains each was responsible for.

On the day of the study, the participants constructed 'Simplifiers' from the timetables to use as job aids. In addition, a TRUST (Train Running System on TOPS [Total Operations Processing System]) display and print out were available on the Panel's faceplate during the simulation. The trials were organised into three sessions:

a Practice Trial; Trial 1 - East Holmes, High Street, Pelham Street and West Holmes control areas (with CCTV MCB for the High Street LX and two signallers); Trial 2 - East Holmes, High Street and Pelham Street control area only (with LX pedestal MCB for the High Street LX and two signallers).

The first 15 minutes of the scenario (lowest train activity level) was used as a practice run. A large wall clock was used to co-ordinate the trial and the minutes and half minutes were called out. At the end of the practice, each participant made Bedford Workload Ratings separately. They then made ratings on the 6 NASA-TLX scales, using a computer, and then went through the paired comparisons procedure to derive weightings for each of the scales. These weightings were used to produce an overall workload score for this practice session and all subsequent TLX ratings.

The first trial session was for two signallers, one controlling West Holmes area, and one the remaining area. After every 15 minutes a short break was made to collect Bedford Workload Ratings and TLX ratings.

The second trial session used only the second, busier half of the scenario due to time constraints. Workload ratings were collected after 15 minutes and 30 minutes, which corresponded to 45 minutes and 60 minutes in the first trial.

5.3 Results

Workload measures were made by all the participants, including the observer who gave estimates of workload for each period.

All the participants confirmed that 2 signallers for both configurations under test could operate the Signal Box. With the West Holmes control area included, the participants substantiated the need for the High Street LX to operate using CCTV controlled from the NX panel to support the staffing complement of two. Having two signallers and a separate Crossing Keeper was likely to cause errors and lead to train and road traffic delays.

Table 4. Workload ratings for the three sessions

| | No. of trains | NASA TLX Ratings | | | Bedford Ratings | | |
|-----------------|---------------|--------------------------------------|----------------------------|----------|--------------------------------------|----------------------------|----------|
| | | Signaller 1 (West Holmes & MCB CCTV) | Signaller 2 (Lincoln Area) | Observer | Signaller 1 (West Holmes & MCB CCTV) | Signaller 2 (Lincoln Area) | Observer |
| Practice | 3 | 24 | 48 | 49 | 3 | 3 | 3 |
| Trial 1 | | | | | | | |
| 15 mins | 3 | 20 | 39 | 46 | 3 | 3 | 3 |
| 30 mins | 4 | 50 | 72 | 53 | 3 | 5 | 3 |
| 45 mins | 5 | 55 | 82 | 67 | 7 | 7 | 4 |
| 60 mins | 6 | 48 | 82 | 52 | 3 | 7 | 3 and 4* |
| Trial 2 | | Signaller 1 (Crossing Keeper) LX MCB | Signaller 2 (Lincoln Area) | Observer | Signaller 1 (Crossing Keeper) LX MCB | Signaller 2 (Lincoln Area) | Observer |
| 45 mins | 5 | 33 | 68 | 72 | 1 | 7 | 6 |
| 60 mins | 6 | 35 | 65 | 59 | 2 | 7 | 6 |

*3 for Signaller 1, and 4 for Signaller 2

They all expressed concern with the low workload and boredom that a crossing keeper would experience with a separate pedestal as tested in the second trial.

The trials confirmed that three signallers would not be practical to operate the High Street SB should the West Holmes control area be incorporated. The High Street LX would need to be controlled using CCTV to support a staffing complement of two signallers. The trials also confirmed that without the West Holmes control area incorporated into the High Street SB, there would be good reasons to convert the High Street LX to a CCTV operated from the panel.

The trials endorsed the views of the SMEs that two signallers could control the Signal Box with or without the addition of the West Holmes control area. If West Holmes is not included, there should be some duplication of the West Holmes fringe to give advanced notice of approaching trains.

The study showed that a peak of 6 trains in 15 minutes is acceptable if the total for the 45 minutes following does not exceed 10 trains (16 in the hour). It was recommended that workload should be reassessed if more than 16 trains per hour, or 6 in 15 minutes are scheduled.

5.4 Study 2 Discussion

General Findings. The scenario can be seen as a worst case in that it represents a period of much higher activity than currently in the area. However, an increase in train movements in the area is expected resulting from increased traffic from the Humber International Terminal at Immingham (e.g. coal imports).

The Workload Ratings from the trial indicated a level of activity that is unsustainable at the levels reported for the Lincoln area excluding West Holmes between 13:30 and 14:00 for both trials, and possibly for the Lincoln area including

West Holmes area for the 13:30 to 13:45 period. The main sources of the high ratings on the NASA-TLX were mental demand and temporal demand.

With most failures, e.g., a barrier failure, there would be a high communication and cognitive workload dealing with the failure, communicating with other persons to remedy the problem and managing the train movements at the point of failure. The volume of train movements would reduce, and in a worst-case scenario, would stop all together. During a failure the telephone concentrator and telephones would be extensively used.

Practice Effects. The participants did not know the schedule, as it was a hypothetical timetable loaded with more trains to test high workloads. It is noted that the TLX workload ratings show a drop between the Practice and the first Trial 15 minutes, which is a repeat of the Practice session. In the second trial the signaller has to control the High Street LX via the crossing keeper, but otherwise this task is the same. Controlling the crossing and ascertaining its position via communications with the crossing keeper was thought to be at least as demanding as controlling the crossing by CCTV. However, the workload levels were lower, which was probably because of a practice effect.

Simulation fidelity. The NX Panel was represented in good detail. However, the top and bottom of the layout were reversed so that the directions for the branching lines were the opposite of those expected, and the Up and Down lines were reversed (due to the status of the current panel drawing). Movement of the train positions every minute meant that some track sections were jumped, and it was more difficult to anticipate when trains would be at a particular point (e.g. level crossing, entry or exit point). Route setting was simulated, but the routes were not highlighted due to the constraints of the mock-up. There were some conversations; between the signallers, the ergonomics study team and other observers, together with the time announcements that would not have been expected in an operational environment. This could have added to the signallers' workloads.

Is Workload is too high? The number of train movements in the scenario was higher than that currently experienced in the Lincoln area. It is possible that the workload in the scenario is too high to be sustained. Even if this is the case, it is acceptable for short durations interspersed with lower levels of intensity. However, it is more likely that the factors examined above are at least partially responsible for the elevated workload levels. All the participants agreed that the proposed signalling area was workable by two signallers.

5.5 Study 2 Conclusions

The workload ratings are likely to reduce with practice, familiarity with the timetable, and the greater feedback afforded by the real equipment. The simulation demonstrated that a small increase in train movements could have a huge effect on the signaller's workload. Any increase in throughput planned for the area should take account of this.

The inclusion of West Holmes area has not been shown to reduce workload for the signaller controlling the High Street area in this scenario. The previous study presented Bedford Scale estimates for workload for a 15 minute scenario with 6 trains and two signallers. These ratings were 6/7 for the Lincoln control area, and 4 if this included West Holmes. The Observer estimated the workload to be the same as in the previous study, that is lower with West Holmes included (6 versus 4). However, for the signaller concerned the workload was indistinguishable on the Bedford Scale, but lower without West Holmes on the TLX, which may be due to a practice effect. The flexibility offered by control of the slow lines still make it desirable that the West Holmes control area is controlled from the same Signal Box.

The participants agreed that the Signal Box should be staffed by two signallers and is manageable on that basis. The use of a separate Crossing Keeper was not favoured because of the potential problems that could arise with very low workload levels and boredom. These problems may include delays to both trains and road traffic owing to inefficient use of the barriers.

The maximum sustainable workload level demonstrated was 4 trains in 15 minutes. A peak of 6 trains in 15 minutes is acceptable if the total for the 45 minutes following does not exceed 10 trains (16 in the hour). It should be stressed that these conclusions are very conservative. The acceptability of up to 16 trains an hour under the poor conditions of the study (from a workload perspective) means the true maximums are likely to be higher. Further studies would be required to extend these limits if train throughput is likely to exceed them.

The NASA-TLX raw scores and weightings show that temporal demands and mental demands are the major contributors to the levels of workload experienced in this study. Giving early warning of trains entering the control area from West Holmes should reduce these temporal demands.

6 Discussion

The methodology employed showed a number of key benefits. The early desktop scenario run throughs with the Bedford workload scale was useful. It helped eliminating some design options, such as one signaller controlling the whole of the new area. It also introduced the concept of mental workload assessment to the stakeholders. Measures made at these early stages showed consistency with those obtained with a full scale mock up. However, this was not an academic study and the numbers involved and the experimental design were such that no statistical conclusions could be extracted. The combination of two workload assessment techniques draws on the strength of both. The Bedford scale is quick and easy to administer, while the TLX offers more rigor and some diagnostic capability.

The use of a full scale cardboard model together with the realistic scenarios enabled a very quick assessment of four different panel layouts as well as the positioning of other equipment in the proposed new Signal box. It is envisaged that it is a user centred design approach that could be applied across a range of industries as well as rail.

Since the study was completed huge advances have been made in Railway Ergonomics. These include developments of specific ergonomic tools for the rail industry and advances in signal control centres.

The University of Nottingham and Network Rail have produced a suite of tools for the assessment of signallers' workload. Pickup [15] gives an overview of each. The toolkit comprises: The Integrated Workload Scale (IWS), Operational Demand Evaluation Checklist (ODEC), The Workload Principles, Adapted Subjective Workload Assessment Technique (ASWAT) and the Workload Probe. In addition they have produced an Activity Analysis Tool to help record activities when assessing workload.

Pickup and Wilson [16] report how the repertory grid technique was employed in the construction of ODEC. It is presented in a spreadsheet that encompasses the components of a signal system and provides an output of the proportion of high, medium and low workload factors that are present. The components include: numbers and types of train movements, points, signals, crossings, line speeds and disruptions. Different spreadsheets are applicable to different types of signalling system: Lever frames, NX Panels and IECCs. An ideal system would show roughly equal proportions of high, medium and low components. ODEC would be the initial assessment that the authors would use if carrying out a similar study today. Indeed, both were subsequently trained at Network Rail's Euston office in the methodology.

Pickup et al [3] describe the development of the IWS. The IWS is a uni-dimensional scale of 9 points. The anchors and descriptions for the IWS are similar to those that are in the Bedford Scale, but with less emphasis on spare capacity. Table 5 gives a comparison.

The IWS can be administered using a keypad device, a touch screen, paper and pencil or a device called an Actiwatch. Although the scale is similar to the Bedford scale the fact that it has been developed specifically for the rail industry makes it the preferred tool. This is a fairly non-intrusive method that typically is applied every minute. This would have fitted well with the way the authors ran the scenarios in the case study, with trains being moved at minute intervals.

A-SWAT is a rating scale similar to SWAT but with the Stress scale replaced with Time Pressure; stress can be viewed as a weakness in the signalling world, Pickup [15]. A second adaption is that the time consuming process of scale development for each participant has been removed. Instead equal weighting is given to each of the three scales. This overcomes a barrier that prevented the authors selecting SWAT in the 2002 assessments. A-SWAT makes a useful addition to the present day toolkit.

Table 5. Comparison of IWS and Bedford Scales

| IWS Rating | IWS Description | Bedford Choices | Bedford Rating | Bedford Description |
|-----------------------|--|--|-----------------------|--|
| Not Demanding | Work is not demanding at all | Workload satisfactory without reduction. | 1 | Workload insignificant |
| Minimal effort | Minimal effort required to keep on top of situation | | 2 | Workload low |
| Some spare time | Active with some spare time to complete less essential jobs | | 3 | Enough spare capacity for all additional tasks |
| Moderate Effort | Work demanding but manageable with moderate effort | Workload not satisfactory without reduction. | 4 | Insufficient spare capacity for easy attention for additional tasks |
| Moderate Pressure | Moderate pressure, work is manageable | | 5 | Reduced spare capacity. Additional tasks cannot be given the desired amount of attention |
| Very Busy | Very busy but still able to do job | | 6 | Little spare capacity. Level of effort allows little attention to additional tasks. |
| Extreme Effort | Extreme effort and concentration necessary to ensure everything gets done | Workload not tolerable for the task | 7 | Very little spare capacity but maintenance of effort on the primary task not in question. |
| Struggling to Keep Up | Very high level of effort and demand, struggling to keep up with everything | | 8 | Very high workload with almost no spare capacity. Difficult to maintain level of effort. |
| Work too Demanding | Work too demanding – complex or multiple problems to deal with and even very high levels of effort is unmanageable | | 9 | Extremely high workload. No spare capacity. Serious doubts on ability to maintain level of effort. |
| | | Not possible to complete the task | 10 | Task abandoned. Operator unable to apply sufficient effort. |

Garner et al [17] developed a system called ATLAS to model train driver's workload. It was based on activities derived from HTA and modelled as VACP demands. The scenarios used were based on system not using ARS (similar to NX operation). Predictions of workload were compared to signaller's ratings using IWS at minute intervals. It was found to have high correlations between predicted and actual visual, auditory and psychomotor activities, but poor for cognitive predictions. The advantage of a model based approach is that new designs and timetables can be evaluated at the early stages of development.

Development and expansion of the IECCs for railway signalling has changed the tasks for most modern signallers. The lever frames and NX Panels are being phased out. The IECCs have Automated Route Setting (ARS) that reduce workload or allow a signaller to administer a larger area. ARS works by using the timetable to resolve conflicts (Balfe et al [18]). During periods of perturbations the signaller may have to intervene to produce an efficient solution for routing trains. The ARS can run in three modes: fully automatic, auto-route functionality (low automation) and fully manual. The manual mode is in effect the same as an NX panel but with a computer screen

interface rather than a track diagram with physical buttons. IECCs all have built in simulation modes so they offer the highest fidelity representations that could be used for workload studies and training.

The two studies were limited in their assessment of workload for the new SB. This was primarily in the duration of the scenarios and the small number of participants. However, the process of designing and conducting the assessments together with other activities of the study (HTA, Error Analysis, interviews, observations and layout assessments) led to a participatory design process. The relevant stakeholders included the signallers, SMEs, the engineering design firm, and the management. The output was a physical design and set of recommendations that had evolved through the duration of the study. Garner et al [17] state that although it is possible to examine workload using simulators (e.g. IECCs) it is less practicable for NX panels as simulators are costly to build. The use of the cardboard mock up as a quick, low cost, low fidelity simulation provides an alternative solution. It was pivotal in this design process and we recommend its use in similar projects.

References

1. Millen, L, Sharples, S.C., Golightly, D., Balfe, N. (2009). The application of cognitive work analysis in rail. Proceeding of Third International Conference on Rail Human Factors. Lille, France
2. Embrey, D.E. (1986) SHERPA: A systematic human error reduction and prediction approach. Paper presented at the International Meeting on Advances in Nuclear Power Systems, Knoxville, Tennessee.
3. Pickup, L., Wilson, J.R., Norris, B.J., Mitchell, L. and Morrisroe, G. (2005). The Integrated Workload Scale (IWS): A new self report tool to assess railway signaller workload. *Applied Ergonomics*, 36 (6), 681-693.
4. Meshkati, N., Hancock, P. and Rahimi, M. (1992). Techniques in mental workload assessment. In J. Wilson and E. Corlett (Eds.), *Evaluation of Human Work. A Practical Ergonomics Methodology*. London: Taylor & Francis.
5. Sauer, J. (2000) Prospective memory: a secondary task with promise. *Applied Ergonomics*, 31 (2), 131-137.
6. McCracken, J.H., and Aldrich, T.B. (1984). Analysis of Selected LHX Mission Functions: Implications for Operator Workload and System Automation Goals, U.S. Army Research Institute, Fort Rucker, AL, USA.
7. Farmer, E.W., Belyavin, A.J., Jordan, C.S., Bunting, A.J., Tattersall, A. J. and Jones, D.M. (1995). Predictive workload assessment, DRA, Farnborough, UK. Final report DRA/AS/MMVCR95100/1.
8. Dahl, S, Laughery, K.R., Hahler, B., Lockett, J., and Thein, B., "CREWCUT - A computer modeling tool for studying dynamic human performance under conditions of high workload." Proceedings of the International Ergonomics Association Meeting held in Paris, France, July 1991.
9. Cooper, G. E. and Harper, R. P. (1969). The use of pilot rating in the evaluation of aircraft handling qualities. Technical Report TN D-5153, NASA.
10. Reid, G.B. and Nygren, T.E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In P.A. Hancock and N. Meshkati (Eds.) *Human Mental Workload*. Amsterdam:Elsevier.

11. Hart, S. G. and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.), *Human mental workload*. North-Holland, Amsterdam. Pp. 139-183.
12. Byers, J.C., Bittner, A. C. and Hill, S.G. (1989). Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary? In A. Mital (ed.) *Advances in Industrial Ergonomics and Safety* (vol. 1). London: Taylor & Francis.
13. Rubio, S., Martín J, and Díaz E., (2000) Multidimensional assessment of subjective mental workload: comparing different ways to obtain an overall workload score. In 1st International Conference on Risk Assessment Tenerife, Spain.
14. Rosco, A.H. and Ellis, G.A. (1990). A subjective rating scale for assessing pilot workload in flight: A decade of practical use. RAE Technical Report 90019. Royal Aerospace Establishment, Bedford, UK
15. Pickup, L. (2007). An Overview of the Development, Testing and Application of the Workload Toolkit. Report No. IOE/ RAIL/06/02R IOE, University of Nottingham, UK.
16. Pickup, L. and Wilson, J.R. (2007). Mental workload assessment and the development of the operational demand evaluation checklist (ODEC) for signallers. In J.R. Wilson, B Norris, T Clarke and A Mills (Eds.), *People and Rail Systems*. Ashgate, Aldershot, England. Pp. 215-213.
17. Garner, T., Newman, M., Lowe, C., and Hamilton, W.I. (2007). Prediction of signaller workload. In J.R. Wilson, B Norris, T Clarke and A Mills (Eds.), *People and Rail Systems*. Ashgate, Aldershot, England. Pp. 237-245.
18. Balfe, N., Wilson, J.R., Sharples, S., and Clarke, T. (2009) Effects of level of signalling automation on workload and performance. *Proceeding of Third International Conference on Rail Human Factors*. Lille, France