The Integrated Workload Scale (IWS): A new self-report tool to assess railway signaller workload

Laura Pickup\textsuperscript{a}, John R Wilson\textsuperscript{a,}\textsuperscript{*}, Beverley J Norris\textsuperscript{a}, Lucy Mitchell\textsuperscript{b}, Ged Morrisroe\textsuperscript{c}

\textsuperscript{a}Centre for Rail Human Factors, Institute for Occupational Ergonomics, University of Nottingham, Nottingham NG7 2RD, UK
\textsuperscript{b}CCD Design & Ergonomics Ltd, Golden Cross House, 8 Duncannon Street, London WC2 N4JF, UK
\textsuperscript{c}Network Rail, 40 Melton Street, Easton, London NW1 2EE, UK

Received 25 October 2004; accepted 31 May 2005

Abstract

Network Rail, who own the railway infrastructure in the UK, have been interested in the assessment of mental workload (MWL) of signallers and control staff for some years. A new model of MWL has been proposed within which to develop a suite of new MWL analytical and empirical assessment tools. One of these is the Integrated Workload Scale (IWS), developed and tested for signallers. This paper describes the development of the IWS and its subsequent testing within two full-scale simulator trials with an NX (entry/exit) panel and an Integrated Electrical Control Centre (IECC) system, and then in the field at signal boxes. The IWS has proven to be a valuable measure of individually experienced peaks and troughs in workload over a period of time or within a particular set of scenarios. It is acceptable to signallers and maps well onto assessments of expected workload on the basis of timeline analysis and subject matter expert commentaries.

\textcopyright 2005 Elsevier Ltd. All rights reserved.

1. Introduction

To say that mental workload (MWL) 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. For drivers there have been some reports of psychophysiological measurements of workload (e.g., Caban et al., 1993; Myrtek et al., 1994). Recently, Hamilton and Clarke (2005) developed and proposed ATLAS as an analytical tool to assess train driver workload, from an information processing angle. For signalling, there have been one or two contributions. Reid et al. (2000) attempted to predict railway signaller workload with technical systems change by using task and timetable analysis and the NASA TLX workload assessment tool (Hart and Staveland, 1988); McDonald (2001) started with NASA TLX and adapted this to make it more suitable for the work of rail network controllers.

A number of techniques were employed in our own early signalling workload studies (see Bristol and Nichols, 2001; Nichols et al., 2001; Wilson et al., 2001). The full NASA-TLX was used as well as a cut down version of three of its scales—“planning and deciding”, “time available” and “physically tired”. The US Air Force Flight Test Centre (AFFTC) seven-point scale (Ames and George, 1993) is a single rating of workload combining perceived demands from activity level, the system, time and safety concerns (see O’Brien and Charlton, 1996) and was adapted by us for use in the railway. A measure of work underload, with items for “under-challenged”, “disinterested”, “bored”, “difficult to concentrate”, etc., was developed. In a different approach, static and dynamic load measures of environmental variables were related to subsequent analysis
of video recorded activities and self-report ratings, as a start in producing some parameters for what constitutes loading factors.

After this early work Network Rail came back with a number of questions such as: can we measure peaks and troughs of the input load on signallers and their (perceived) efforts to meet this load? Can we distinguish high and low loading signalling sites or situations? What types of tool can be developed for prediction of workload at new sites or jobs? As a result, a suite of new or adapted tools for use in rail signalling workload assessment has been defined and developed (see Pickup and Wilson, 2004 for full details). A suite of assessment tools was the most effective way of assessing the number of relevant dimensions; a single tool is not capable of assessing all variables relevant to the signaller (Brookhuis and De Waard, 2002; Williges and Wierwille, 1979). Within this suite of tools, an Integrated Workload Scale (IWS) of nine points has been produced, inspired by the Instantaneous Self-Assessment (ISA—Brennan, 1992) Scale. The IWS is the focus of this paper, which examines its basis and development and then trials of its use in simulators and in signal boxes. For information about the other empirical and analytical tools in the suite, see Pickup and Wilson (2004) and Pickup et al. (2006).

2. Nature of MWL and its measurement

One of the most widely debated notions in ergonomics/human factors is that of MWL. It is a seductive concept, apparently meaningful to specialist and layperson alike, which can be adapted to fit many contexts. As a result, MWL—or often just workload—is used to denote quite different factors or situations, and has a plethora of measurement methods associated with it. Loose and non-operationallised use of the term, and consequent variations in how MWL is assessed, can leave clients of ergonomics bemused. Another difficulty is use of a MWL measure in a particular setting when it was created for an entirely different purpose in entirely different circumstances.

There are no universally accepted definitions for MWL, nor is there agreement on any one measurement approach or tool: “There has been much discussion of an acceptable definition of MWL . . . . As with other areas of ergonomics this has proven difficult...consider the wide array of definitions of MWL [in] the first significant collection of papers... (Moray, 1979)...and things have not got that much better...The problem is the proverbial one of the chicken and the egg... [are we] trying to use our [imperfect] knowledge of psychological processes to derive measures of MWL... or using MWL measures to investigate...psychological processes...A bit of both” (Megaw, 2005).

Over a period of some years, the authors have been called in to make judgements and provide insights about MWL in rail signalling, on the basis of concerns such as:

- What are the number, complexity and interaction of tasks performed by someone, over a period of time or at one point in time?
- What are the requirements of the tasks to be performed in a certain time compared to the capacity of the individual(s) to complete them successfully?
- What load or effort is perceived as experienced by someone, over a period of time or at one point in time?
- What combination of functions can be performed successfully by a person or a team in different situations and scenarios?
- How compatible are work systems such as computer interfaces, job designs or procedures with the functions that need to be completed, in terms of their influence on someone’s performance and well-being?

These concerns suggest a major distinction between workload as imposed by the system, somehow measured independently of any individual, and workload as rated by those individuals carrying out the “work” that is “loaded” on them, whether this reflects perceived task difficulty or level of effort exerted etc (see Hart and Staveland, 1988). In the particular context of rail, it seems that MWL is seen by industrial stakeholders in three broad ways. In the first view it is seen in terms of “the work that loads” (or the work that creates demand); this leads to questions such as how many functions can one or more signallers handle (under normal and perturbed conditions), will a new technology in the signal box or train cab increase the demands on the operator beyond their capacity to cope safely and effectively, or how will design and sequence of information inputs (signs or signals) load a train driver? Within this view, there is need for an analytical tool or an expert assessment method, which can identify levels and limits to the “work that loads”, and considers the actual demands of the work in the way it is organised. This will also allow understanding of the contribution of socio-technical factors and time pressures to workload, and measures may be related to information processing theories such as multiple resource theory (Wickens, 2002).

In a second broad view, MWL is seen as “the experience of load” by the operator. In this case the rail industry need is for workload to be measurable through operator reports and ratings. Such reports may be explicitly or implicitly of different factors: the many external and internal sources of load; the different personal experiences of demand; the level of effort the operator is prepared to, actually does, and perceives that
they put in; and the effects on them and on the system of the demands and of their efforts.

In a third, related, view MWL is seen more as an actual decrement in performance, as the measurable effects of changes in operator behaviour and their influence on the system. In this case we would infer MWL from changes in measured performance. But this is fraught with difficulty. As task demands increase there are not necessarily associated increases in operator self-reported load nor decreases in performance, for many reasons. MWL is influenced by operator perceptions of the task and by skill, experience and training as well as by task demands. Dissociation between workload and performance frequently occurs. Operators may modify their workload on the basis of explicit or implicit feedback, changing strategies or motivation. Performance outcomes themselves can modify the tasks. Different types of performance at different times may necessitate different MWL measurement approaches: differences which might be influential are use of controlled or automatic processes, data-limited or resource-limited performance, use of single or multiple resources, and task management which includes task switching—an aspect of efficient performance that may bring its own load.

To make some sense of the different conceptions of MWL we have produced a model embracing several of the important ideas, including imposed, internal and perceived load, demands, effort and effects (Pickup et al., 2006). This was originally for rail signalling, and was applied subsequently to underpin research on train driving MWL. In parallel we have clearly proposed that the different MWL notions and the different rail industry needs will necessitate a suite of MWL tools which together will build up a total picture of workload for a function, system or scenario. Large numbers of workload measures exist, until now none produced directly for the rail industry, and a number of them have shown problems with appropriateness of wording and form. The majority of the methods have been developed for a North American military or aerospace population, usually in simulators or else highly physically constrained settings, and may not transfer to best effect in fieldwork with a European and civilian workforce. Although we had no desire to develop new workload assessment tools if none were needed, we were not convinced that this was the case.

Following Megaw’s (2005) use of Hill’s (1987) framework, we can divide MWL measurement approaches into two broad categories, analytical and empirical. Several different tools and techniques might be used within the approaches, none of which is ideal in all circumstances but all of which may have some merit at times. Within analytical approaches we have checklists and walkthroughs, task analysis (usually with timeline analysis), simulation, and computational methods. Within empirical approaches we have performance measures, psychophysiological methods, and self-reports.

Criteria for MWL tools have emerged from Network Rail and UK Rail Safety and Standards Board requirements to include: support diagnosis and/or prediction; be applied concurrently and/or retrospectively; support direct assessment and be analytical; take an analytical and/or direct measurement approach; be qualitative and/or quantitative in form; permit instantaneous and/or continuous measurement; allow tracking of peaks and troughs of load; be “chronic” (i.e. make assessment of load over time) and/or “acute” (i.e. make assessment at a point in time); and be informative about performance consequences. These are in addition to the normal requirements for validity, reliability, acceptability and sensitivity. No one tool or measure will meet more than a few of these criteria. As tools are defined and used, different levels and types of validation will have to be carried out. Many of the criteria that have been given in fact can be seen as spectra and there is rarely any best or worst point on these spectra. For instance, diagnosis can be at different levels; it need not mean only the identification of causal factors, and a simple workload classification—even probable/possible/no—is a diagnosis of sorts. Likewise prediction does not have to be quantitative and such a measure therefore does not necessarily have to have computational power; for instance we can predict that a new cab system will not increase workload via application of expert assessment checklists or self-report scales in a simulation.

3. Self-reported assessment of workload

Some authors believe that self-report gives the most sensitive and accurate reflection of MWL (Hart and Staveland, 1988). Others suggest that self-reports can reflect the actual effects on performance better than measures of task demands (Jensen et al., 1994), benefiting from the operator’s insight into an increase in effort prior to performance degradation. When an individual is rating themselves they may well be considering how well they are coping, the resources they are using and the amount in reserve, previous experiences and their level of motivation (Muckler and Seven, 1992).

Self-report workload measures fall into two categories—multidimensional and unidimensional workload scales. Multidimensional scales explicitly represent the dimensions of workload and allow a rating to be obtained for each dimension, e.g. NASA TLX (Hart and Staveland, 1988), or Subjective Workload Assessment questionnaire (SWAT) (Reid and Nygren, 1988). Unidimensional scales represent the concept of workload as one continuum—e.g. the Bedford Scale (Roscoe,
represent the concept of workload in the most intuitive way to signallers, thus increasing validity of the measure. Cainesmith and Kendall (1963) consider dimensions defined in terminology suggested by end-users to provide a higher degree of content validity. Furthermore, Brennan (1992) concluded from experiments during the development of descriptors for ISA that a descriptor set that was more reflective about how people feel about their workload levels may prove to have greater meaning and suggested this could be a “promising pointer for future investigations”. We decided to develop a unidimensional scale specific to rail signalling that could achieve the practicality and simplicity of ISA but used descriptors that reflect how signallers themselves feel about their workload levels, in an absolute graphic rating scale.

Construct validity of a workload scale can be enhanced when the descriptors used consider workload theories (Nygren, 1991) and psychometric theories of measurement (O'Donnell and Eggemeier, 1986; Kantowitz, 1992). The psychometric properties of any workload tool are critical for it to measure the relevant dimensions of workload in the context required (Nygren, 1991). By basing tool development on relevant theory, and on the perceptions and language used by the signallers, we aimed to enhance the validity—face, content and construct—of our IWS.

4. Development of the IWS

Psychophysical scaling as a part of psychometric measurement provides a powerful measurement approach to quantify the subjective experience of workload (Gopher and Braune, 1984). Thurstone’s method of equal-appearing intervals (Thurstone, 1927; Oppenheim, 1992) was adopted, balancing a robust development process of a scale of equal appearing intervals with efficiency in its application. The conceptual framework and theories of workload provided a set of dimensions as the core psychometric properties to be reflected within the workload measure, to ensure construct validity and produce a tool capable of capturing workload. These main dimensions were load, demand, effort and effects.

Terms more specific to the signaller’s workload dimensions were collected through interviews and observations as a part of workload investigations at a sample of signal boxes. The interview data were coded based on the conceptual framework (Pickup et al., 2006). Systematic analysis of interview data extracted perceived sources of load, causes of effort and demand and types of effects, when the signaller performs the key functions of gaining awareness of the situation, making decisions and planning strategies, and acting to implement plans. This provided key workload dimensions in terminology familiar to signallers. Examples of terms
and phrases used are: load—amount of work, jobs, tasks, situations, responsibilities, problems, time available; demand and effort—concentration, focus of attention, busy, effort, demanding; and effect—pressure (time and individual), frustration, struggling, spare time, managing. It was rare that only one term or dimension was used by any one signaller to describe the workload they experienced. This supported the view of Ames and George (1993) that raters using the AFFTC are capable of integrating dimensions in interpreting the term workload, and in fact this appears the more natural approach to expressing perceptions of workload experienced. Therefore descriptors were developed to present dimensions in combination and were included in the pool to be judged by signallers as to relevance for describing their workload.

Typical phrases transcribed from the interviews were “struggling to keep up” and “feeling under pressure”. The main aim was to ensure that the descriptors were in a language familiar to signallers and reflected the most relevant workload dimensions. We needed to provide good rating scale cues to allow sufficient anchor points along the workload continuum in order to facilitate the signallers’ judgment (Guildford, 1954), and these were selected so that the final scale is unidimensional as a continuum to represent the overall variable of workload.

Two rail human factors researchers independently agreed that a final set of 47 descriptors had clarity and relevance to signaller workload, and that this number was manageable when considered by signallers over a short period of time. It was not possible to gather a large number of judges to complete a card sorting activity. In a minor adaptation of the Thurstone technique, a questionnaire was designed to administer the card sort. The descriptors were numbered and presented in random order and signallers judged each statement for whether the term appropriately described high or low levels of workload. The workload continuum was represented by a line from high to low workload and rail signallers were asked to put each descriptor in one of 11 boxes that represented equal points along the continuum. An additional box was titled ‘confusing’ for any obscure or confusing statements. The questionnaire was piloted with signalling subject matter experts (SMEs) and human factors experts, and 130 questionnaires were then distributed by hand to signallers, to allow a verbal explanation to be provided and encouragement given to complete the questionnaire. The locations were chosen as representative of the main generations of signalling systems, that is, lever frame signal boxes, NX-panel systems (entry–exit) and an Integrated Electronic Control Centre (IECC) system. Response rate was 23% (30 out of 130), not good but sufficient (a minimum of 20 judges has been suggested by Dane (1990)).

Two judges who placed more than one-third of their responses in the same box were excluded, so the final analysis was completed with 28 judges. The next stage was to consider the spread of judgments made on each statement; the wider the spread the more ambiguous the statement. The median and semi-interquartile-range were calculated for each statement; where there were two closely positioned phrases the mean was consulted to assist in the final decision. A frequency distribution was calculated for the phrases collected in each box. A cumulative-percentage frequency graph was produced (Fig. 1), the 11 box categories form the X-axis and the percentage frequencies the Y-axis, and the graph suggests the data collected were relatively well distributed.

The optimum number of points for a unidimensional scale is debatable. With too few points a scale will lack sensitivity and with too many the rater may be unable to discriminate to such a fine level. The issue of central tendency highlighted earlier with the ISA tool was a good indication to develop a scale with a greater number of descriptors. Referenced in Guildford (1954) analysed over 23,000 ratings and suggested that a nine-point scale was optimal for bipolar scales. Freeman and Freeman (2000) suggest that scales with between 7 and 12 steps provide the maximum discrimination. Therefore the IWS was set to have nine descriptors, with the intention to assess scale length and a possible reduction in this as part of the subsequent validation process.

There was a strong consensus on the descriptors even at the extreme points and so data across the range were included. To move from 11 categories to a nine-point scale, a ratio of 1.25 (10 intersections for 8 intervals) was applied. Therefore the descriptors were reviewed to extract those that had medians closest to 1, 2.25, 3.5, 4.75, 6, 7.25, 8.5, 9.75 and 11. This is presented as the final IWS in Fig. 2.

For greater usability in the presentation of the scale, the key workload dimension within each descriptor was highlighted and presented to the left of the full descriptor. At a glance the key dimensions can act as
prompts to a signaller familiar with the scale from prior use, with the scale presented as colours rather than numbers to subtly stress rating by description. The output from the IWS tool is a graphical trace of workload experienced by the operator during a particular scenario, task, event or period within a shift (Fig. 3). Although a mean workload value over time could be calculated (but only with caution), the greatest use of the IWS is to track the peaks and troughs either in a simulator or field environment at intervals to suit needs (e.g. every 1, 2, 5, 15 min, etc.).

The scale is packaged to allow several modes of presentation. The simplest for the investigator is a paper chart, in colour, presented in a position that is visible yet not obtrusive to the signaller nor interfering with their system of control. The investigator uses a stopwatch to time and prompt the signaller at agreed rating intervals, and then records the rating on paper. The advantages of this approach are its simplicity and no problems of obstructive or malfunctioning technology. The disadvantages are that to create the graphical output the data have to be entered into an excel spreadsheet at a later date, and of course the investigator must be present and concentrating throughout.

Electronic devices were considered in order to streamline the data collection, analysis and presentation processes whilst minimising intrusiveness in the signaller’s environment. Technology that involved any form of radio waves (e.g. wireless or blue tooth) was dismissed due to the potential for interference with control equipment. Equally voice activated devices were considered unsuitable due to surrounding noise levels and the need to transfer between multiple users. A touchscreen laptop computer was used, with a Visual Basic software programme to prompt the signaller at pre-set intervals. An audible alarm (distinct from other control system alarms) alerts the signaller to touch the colour button next to the descriptor most applicable at that time. The software records the rating made, the prompt time and the time the rating was actually received. Two graphs are instantly produced at the end of the session to represent prompt time elapsed and time of actual scoring.

Due to space being limited within some signal boxes a further interface was designed, based on that used to administer the ISA workload tool by Eurocontrol (Hering and Coatleven, 1996). This is a nine-point keypad that can be positioned as desired by the signaller, attached via a port into a standard laptop computer. The keypad provides an audible alert local to the signaller’s position and when the keypad is pressed the rating is recorded using the same software as described previously. Also the Actiwatch has an interface similar to a watch that requires a button to be

Fig. 2. IWS for rail signallers (colour codes run from blue for not demanding to red for work too demanding).

Fig. 3. A typical trace of output from the IWS over a 35 min period.
pressed a set number of times to reflect the rating descriptor being recorded. The data are then downloaded into an excel spreadsheet. The latest software version, IWS 3.0, has been programmed to prompt at each time interval irrespective of a previous rating being entered or not. The signaller is provided with a detailed briefing before starting a trial, to emphasise that supplying a rating is their lowest priority. The IWS tool is capable of highlighting fluctuations in workload, but does not record the reasons for the fluctuations. Video and observation by human factors experts or SMEs can provide valuable information to support and explore the reasons behind workload fluctuations.

5. Simulator trials of the IWS

5.1. IECC trial

A trial was run in April 2003 using the York IECC simulator with scenarios developed by Network Rail Operational Experts (see Fig. 4). There were five facilitators: two Network Rail operational experts, one simulator facilitator responsible for running the scenarios, and two human factors experts—responsible for IWS recording, timeline recording and trial administration.

Seven participants took part, taken from the roster on duty in the York IECC on trial days with ages from 24 to 50 years (mean 39 years) and experience at York IECC from 3 months to 14 years (mean 4.5 years). All participants were familiar with York IECC and the simulated areas of control presented in scenarios. Participants were told they could take breaks between scenarios or leave at any time if they wished and they would not be penalised in any way. They were given the opportunity to ask any questions before the trial began. They were further informed that the results of the trial would be used for the purpose of the study only. Participants remained anonymous and participation did not carry any implications for Network Rail personnel involved.

Realistic scenarios were designed for the simulation with all participants familiar with the area and the geography used. The timetable was not one that the signallers had used before, however, and they were given time to familiarise themselves with it. Although signallers normally worked this area with Automatic Route Setting (ARS), this was switched off for the trial (which can also be the case in practice). Scenario 1 represented ‘normal’ working conditions, the kinds of tasks a signaller expects to perform on a day-to-day basis. Scenario 2 represented higher workload, where the signaller must respond to a perturbation in the system, and was the same as Scenario 1 except that a Hot Axle Box Detector (HABD) was triggered 4 min into Scenario 2, with only a visual cue. In fact, all but one signaller had to be prompted to respond to this alert by a facilitator acting as duty signalling manager. Each scenario was 15 min long.

Signallers were briefed about the purpose of the trial and use of the IWS. Participants were told that they would be alerted with a tone at 2-min intervals, and once prompted they should choose the coloured touch key on the screen associated with the term that best described how they were feeling at that time. It was stressed that this was their lowest priority task and that missed responses were not detrimental to the trial. Participants were given an operational brief providing the context for the scenarios they were about to undertake. They were asked to undertake each scenario in accordance with the rulebook, complete the IWS ratings when required and complete a short questionnaire at the end of the trial.

The scenarios were undertaken both with the IWS and without the IWS. The order of presentation was randomised using a Latin square design to counterbalance learning effects. Throughout the scenario a facilitator recorded signaller activity using the Noldus Observer Tool for a form of timeline analysis. After each scenario the participant was asked to complete a short questionnaire assessing the usability and intrusiveness of the IWS. Finally the participant was asked to retrospectively complete a SWAT questionnaire, which had minor adaptations to the terminology.

The graphs produced from the results show peaks and troughs in workload across the recording intervals (an example is shown in Fig. 5). This also shows timeline analysis data overlaid onto the IWS output. The timeline analysis goes some way to provide a rationale behind the IWS scores.

A positive Spearman Rank correlation was found between IWS and SWAT scores in the ‘normal’ scenario of 0.69 (n = 7, p < 0.05) or abnormal (z = −0.816, n = 7, p > 0.05) respectively. Use of the 2-tailed Wilcoxon signed rank test showed no significant correlation was due to a higher incidence of missed IWS ratings in the degraded scenario. The correlations do suggest that SWAT and IWS are both measuring the same thing (workload).

Participants reported that they found the tool did reflect how hard they felt they were working and gave an accurate representation of their workload. Questionnaire responses showed the IWS was not intrusive, and this was confirmed through the comparison of SWAT data for trials with and without the IWS. Strong correlations were found between SWAT scores with and without the IWS for normal and abnormal scenarios (0.68, 0.97 respectively). Use of the 2-tailed Wilcoxon signed rank test showed no significant difference between SWAT scores with the IWS and those without it for either normal (z = −1.0, n = 7, p > 0.05) or abnormal (z = −1.0, n = 7, p > 0.05)
conditions. If the tool was intrusive is likely that SWAT scores would be higher in scenarios when the IWS was used compared to when it was not.

5.2. NX panel simulator trial

A second trial was conducted in August 2003 using an NX panel at Crewe (Fig. 6). This trial used a scenario that generated higher workload for operators, to test the higher end of the rating scale (in the York trial none of the participants scored above five on the IWS). We also wanted to test again if the tool would be intrusive for the signallers’ work. This trial examined individual participant data in detail, rather than averaging across participants, to allow tracking of IWS scores in response
to the demands of the scenario and the individual decision-making process.

There were five facilitators present at the trial: two NR operational experts (responsible for briefing participants, de-briefing participants and providing the SME commentary); two human factors experts responsible for timeline recording, IWS recording, and trial administration; and one Crewe NX simulator facilitator—familiar with the NX, panel simulator and responsible for briefing participants about the panel, running the simulation, and providing the voice communications input.

Ten participants took part—nine trained signallers and one experienced signalling engineer—from various locations throughout the country. All participants were familiar with NX panel operations although none of them were familiar with the Crewe simulator. Experience ranged from 5 to 35 years (mean 18.5 years) and experience with NX panel operations from 2.5 to 20 years (mean 9 years). Two participants did not complete the trial; they struggled with the scenario from the outset and the NR facilitator decided it was not in their best interests to continue.

All participants were briefed regarding the purpose of the trial and were assured that they would remain anonymous, the data would be used only by the outside researchers, the trial was not a personal assessment and there would be no implications for them as NR employees. Should they feel uncomfortable with the trial at any point and they indicate this to the NR facilitator, the trial would be stopped if they wished.

A realistic scenario was developed by NR operational experts, including a four-platform station, a station siding, and a factory siding. Both sidings required phone calls for shunting in and out. A branch line dictated a slot request via phone from the fringe signal box. A simplified timetable was provided. The scenario lasted approximately 90 min and included periods of both high and low workload. At the start there were lower traffic levels and less frequent train annunciations. Forty minutes into the scenario the workload was increased with higher traffic levels and then there was a points failure approximately 45 min into the scenario, followed by a track circuit failure approximately 60 min into the scenario. The points failure was simulated at the entry into the station sidings, and the track circuit failure was simulated on platform four at the station. Relevant communications were controlled from the observations/control room and were role played by the simulation facilitator. These communications were held as consistent as possible between participants.

On arrival, participants were briefed regarding the trial, operational scenario and context and were familiarised with the panel. They were instructed in the use of the IWS and that they must respond to an audible prompt by selecting the appropriate term to describe their workload on the touch screen. Throughout the scenario the participant was prompted to respond on the IWS Scale every 2 min. Activity was recorded using Noldus Observer and by hand by a NR operational expert to provide a timeline analysis and expert commentary on performance. At the end of the scenario the participant was asked questions regarding intrusiveness and usability of the IWS.

Fig. 7 shows an example of IWS data output. A general pattern was found across participants which indicated the sensitivity of the tool. A peak of activity at the beginning of the trial was related to initial route setting (and, possibly, unfamiliarity), and a rise in scores was found when the traffic increases around 40 min, when points fail at 45 min, and when the track circuit fails at 60 min. The York trial scenario had not elicited scores above five ("Moderate Pressure"), but the more demanding scenario in this trial elicited higher scores and indicated that the tool is sensitive at the higher end as well as within the lower range of values.

The timeline analysis again illustrated that increased scores on the IWS were associated with the level of communications and the frequency with which actions are performed. Also, the SMEs were able to observe the effect that higher workload scores had on task output, their commentary suggesting that performance began to deteriorate for many participants at IWS scores around 5 ("Moderate Pressure"); this was shown in a variety of ways such as increased delays, missed telephone calls, omitted reminder appliances, failure to clear signals, route cancellations, and missed failure indications.

The consistency in pattern of the IWS scores across participants is shown in Fig. 8, which also relates the scores to changing scenario demands.

Participant responses to the questionnaire confirmed the sensitivity of the scale, all participants saying it reflected how hard they were working across the
scenario. Participants found it easy to decide on the appropriate describer for their experience of workload at any given time. Only one participant indicated that there may be an issue with intrusiveness in the field, directly related to busier locations, and it is possible that different hardware implementations may address any concerns.

Most participants missed ratings at times and had to be prompted verbally. Where participants recognised that they had missed scores they attributed this to the fact that they either could not hear the prompt or that they were concentrating so hard that they did not notice it. Increased response times were shown at later points in the scenario, consistently with the higher workload scores. As participants get busier, and their workload increases, they can miss the auditory alert due to concentration on the panel. Also, they are already involved in actions and communications and it takes them longer to fit the IWS response into their work. Some participants reported that the touch screen was not very sensitive. Particular hardware implementations may be more appropriate for some participants and

Fig. 7. Example of data output of NX panel simulator IWS scores with timeline activity descriptions and SME commentaries.

Fig. 8. IWS scores over the time of the simulation for all participants in the NX panel trial.
some environments (paper and pencil, keypad, laptop, actiwatch).

6. Field use of the IWS

All the new MWL tools (see Pickup and Wilson, 2004) have been tested in practice in various signal boxes when the Network Rail Ergonomics National Specialist Team make site visits to assess human factors (Morrisroe, 2005). The timeline and IWS are typically recorded for a 60-min period that coincides with the predicted peak workload. The results from the two tools allow the assessors to see a trace of increases and decreases in reported levels of demand and effort and whether these correlate with the activities represented in the timeline. Generally an operational specialist records the timeline information manually, with a recording sheet divided into minutes to record the activity and give some indication of start and end if it ran to more than 30 s. This duration was often exceeded for telephone calls but rarely for any other signalling task—e.g. route setting or data entry for train description. The timeline also recorded walking around the panel and watching a particular event. The activity was video recorded, but on no occasion have we needed to refer to it to amplify the timeline record. The IWS was prompted every 5-min since in the live signalling situation this interval appeared to give the correct trade-off between number of data points and interrupting the signaller to ask for a rating.

Two methods of capturing the IWS data have been used in the field. First, after briefing the signaller the assessors set up the response options paper sheet next to the signaller’s work area and then prompt (every 5 min) for a verbal rating; for reasons of ease of use, we added a numerical value for each of the points. The second method used is the Actiwatch that can be programmed to prompt at the specified intervals and record the rating that the signaller makes by pressing a button several times until the appropriate number is shown. The scale was inverted so that ratings at times of highest load required less button presses from the signaller. This method has the advantage that the signaller can respond whilst away from their workstation.

In every use so far, the IWS ratings vary over the recording period and give a good indication of relative demand and effort relating to the tasks recorded in the timeline. The respondents appear to adopt a response set whereby for the initial series of ratings they refer to the response descriptions but once they settle into a pattern of rating they appear to vary about their chosen ‘default’ level without referring to the scale descriptions. The nine-point scale does not appear to be unwieldy and all levels of the scale have been recorded at different times (but not by the same person). It may be that the response set effect described above actually makes the scale practical and easy to use.

IWS scores for each signaller are indicative rather than categorical and at the moment we use any range values such as mean, minimum or maximum with caution. We most commonly review IWS profiles with respect to demands imposed by the system. These demands—for instance number of train movements, number of point ends, type of signalling information system interface—are assessed by one of the other tools in the MWL suite, the Operational Demand Evaluation Checklist (ODEC—Pickup et al., 2006). In this way, IWS ratings can be compared across sites with similar profiles of demand.

We need to ascertain whether the workload measure is representative of all the individual signallers and the range of task demands that confront them, is not artificiality manipulated by the respondents, and is reflective of general workload rather than just the period of time over which the assessments were conducted. Each of the assessments addresses and observes a single signaller’s experience of workload, and does not necessarily give a consensus view for the signalling team. On occasions when we have suspected that the signaller studied is not representative of the rest of the staff, we have followed up the visit to meet other staff and validate the findings. An instance of this was observed in a signal box where a highly experienced signaller participated in the assessment, but other staff were relatively new to signalling. The assessors are aware of the possibility that some manipulation of results might take place, especially when staff are already complaining about workload, and this risk has been reduced by having an operational specialist on the assessment team. Any bias may be in the opposite direction and a few signallers have commented, after completing the IWS session, that they find it ‘goes against the grain’ to give ratings at the higher levels of effort. Therefore, there is a risk that signallers will downplay their perceived effort.

Since the assessment rarely exceeds 3 h, there may be some doubt about the generalisability of the results. However in practice the remit of the assessment often directs us to consider a particular period—for example the morning or evening rush hours, the periods of possession management, the combination of two areas of control for the night shift, or requests for line crossing by the public. In the workload assessment report, the prevailing state of the railway during the assessment is always described, and it is emphasised that the conclusions and recommendations are made on the basis of what was observed and recorded on the day. This provides a useful platform to consider variations in demand on the overall workload.
7. Discussion

This paper has described the Integrated Workload Scale, a unidimensional scale constructed and validated explicitly for use with rail signallers. The IWS has been demonstrated in two separate simulator trials to be valid, sensitive, and usable for assessing workload in the rail industry. It does not appear to interfere significantly with the job of the signaller in a simulated environment.

In the simulator trials the IWS and timeline/SME commentary show a combined picture that responds to trends and manipulations in the scenario. The IWS identifies peaks and troughs in workload whilst the timeline gives a degree of diagnosticity, indicating the actions associated with higher workload. The timeline indicates that IWS scores increase in relation to the level of communications and task frequency and switching. There is scope for integrating the IWS, timeline analysis, and SME commentary within one software-based tool which would reduce the number of assessors required and could, potentially, aid analysis.

In actual practice to date the tools appear to generate the necessary information for assessors to determine whether there is a workload problem, be effective in identifying the sources of demand, administered within a half day. Therefore the testing and validation programme has shown the IWS to be of value, valid, and presenting few if any problems in application. It is now an accepted measurement and assessment tool for Network Rail.

Acknowledgements

Most of the work reported here was funded by Network Rail and RSSB. Later funding was from the EPSRC Rail Research, UK. We are grateful to all these bodies for the opportunity to explore these important issues. The contribution of Bob Muffett and John Robinson, to development, testing and refinement of the IWS, has been invaluable.

References


