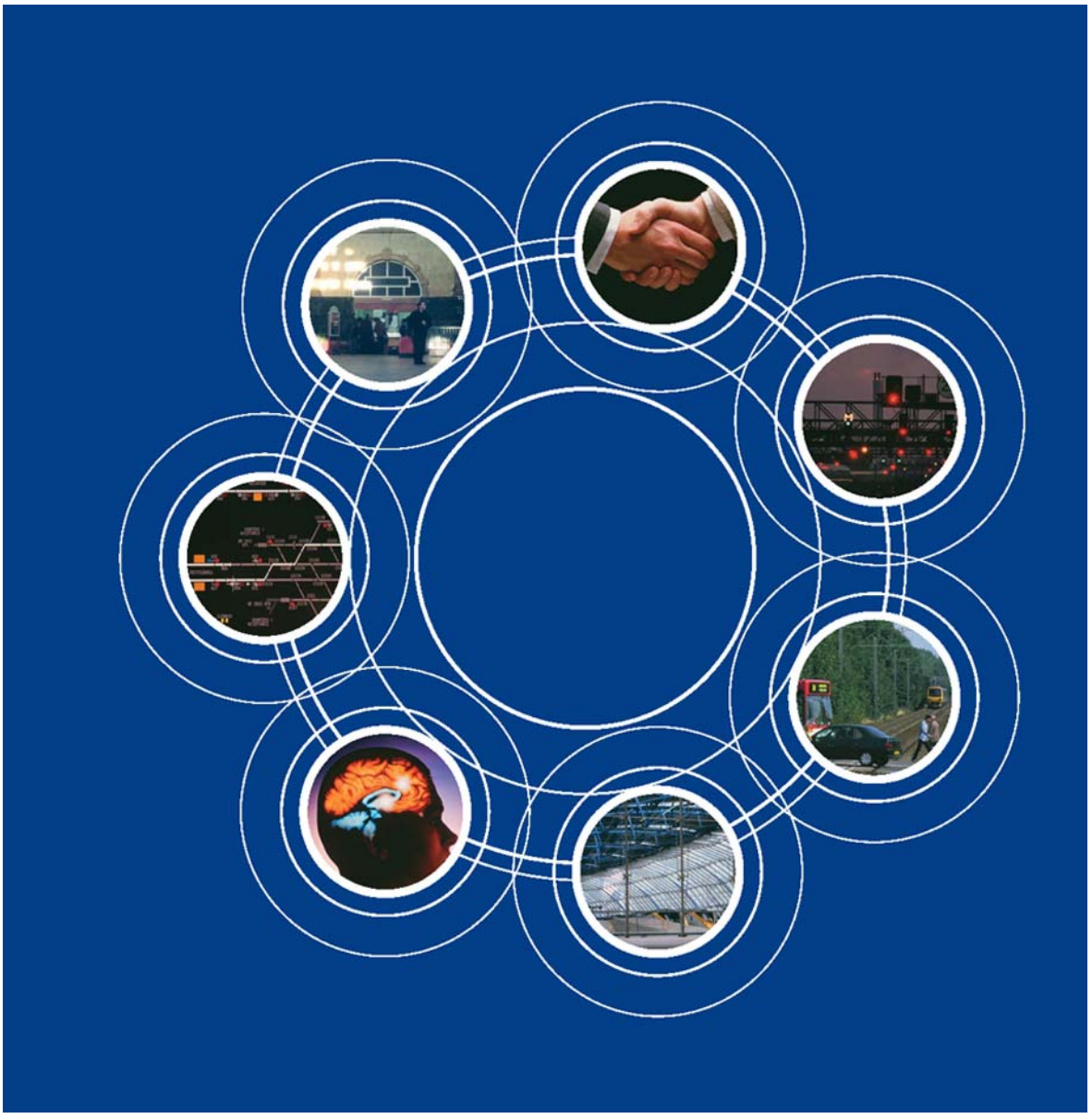
Research Programme

Operations and Management

Driver selection: implementation phase
Final report and technical annexes



Copyright

CONFIDENTIAL

FOR THE USE ONLY OF RSSB MEMBER COMPANIES EMPLOYING TRAIN DRIVERS AND ORGANISATIONS MANAGING THE PSYCHOMETRIC TESTING PROCESS.

There is a publicly available research brief and summary report that do not contain full details of the selection criteria and scoring rules. Some of the information contained in this full report could give an unfair advantage to candidates and it is therefore not to be made publicly available.

Any additional queries can be directed to enquirydesk@rssb.co.uk.

Published: February 2013

Table of Contents

T948: Train Driver Selection: Implementation Phase – Final Report

1 Executive summary

1.1 Background

This report is the main deliverable of project T948 *Train Driver Selection: Implementation Phase.* T948 was the culmination of a programme of research work on train driver psychometric assessment.

The standardised psychometric assessment process consists of selection criteria that are relevant to train driving, an assessment method for each criterion and associated pass marks. Its purpose is to assess whether candidates meet a safe minimum standard in attributes that are necessary for safe train driving. The process that is current at the time of publication of this report (December 2012) is detailed in the Rail Industry Standard RIS-3751-TOM *Train Driver Selection*, Issue One.

At this time, the psychometric assessment process needs to be updated for two main reasons. Firstly, to assess updated selection criteria that are more tailored to modern train driving and meet the requirements of the Train Driver Licence and Certificates Regulations (2010) (TDLCR). Secondly, to address weaknesses in the current assessment process regarding validity and fairness to candidates in ethnic minority groups.

The industry is committed to addressing the limitations of the current process within the constraint of needing a psychometric assessment process that effectively filters out candidates who do not have the aptitude to become safe train drivers.

The overall objectives of industry's work on train driver selection are:

- To implement an updated process for train driver psychometric assessment in RIS-3751-TOM.
- To monitor the effectiveness of this process on an on-going basis so that it can be maintained fit-for-purpose.

RSSB has a remit to assist industry to achieve these aims by making recommendations about what the updated process should be and by undertaking reviews of the process to inform industry monitoring.

In support of this, previous research (project T340 *Psychometric testing – A review of the train driver selection process* (RSSB, 2005) and project T628 *Train Driver Selection – Development Phase* (RSSB, 2010) provided an initial new set of selection criteria, trialled a number of new assessment methods and developed new measures of the behavioural selection criteria. Project T948 was initiated to address outstanding issues.

1.2 Aims

Project T948 aimed to:

- Evaluate the suitability of the proposed methods for measurement of behavioural (non-technical skills) criteria and psychomotor tests for hand coordination and vigilance, in terms of validity, reliability, fairness and practicality.
- Propose a revised assessment centre process that will effectively measure the updated selection criteria, conform to good practice in selection, exclude unsuitable drivers and where possible enable potentially exceptional drivers to be distinguished.

In order to fulfil the latter aim, project T948 drew on evidence from both of the previous projects, T340 and T628.

1.3 Method

The trials conducted during T948 evaluated two tests of vigilance (known as WAFV and VIGIL), a test of hand coordination (2HAND), the Situational Judgement Exercise (SJE) and Multimodal Interview (MMI) for the measurement of the behavioural selection criteria and the Written Communication Test (WCT). These tests are defined in the glossary – (Section 12). The evaluation covered validity, reliability, fairness and practicality. The evaluation covered validity, reliability, fairness and practicality (see Annex 3 section 3 for a full description of the evaluation criteria).

A sample of 146 participants (drivers, trainee drivers and failed candidates) from 17 different operating companies took part in the evaluation trials. Participants completed the assessment methods to provide assessment method score data. Demographic data were collected to evaluate fairness. Train drivers and trainees' managers were approached for information regarding the participants' performance at work. The main data analysis focused on criterion validity and explored the correlations between assessment method scores, operational performance measures and manager ratings of behaviour. Participant and assessor views were recorded to assess face validity and practicality. Information from the test manuals and other evaluation studies of the assessment methods was also collected and taken into account as part of the evaluation of validity and reliability. The full method of the research is described in Annex 3 Section 4.

1.4 Findings

The final recommended set of selection criteria are summarised in section 5 and detailed with full definitions in Table 16. The criteria cover the key skills required for the modern train driving role, satisfy the requirements of TDLCR and reflect the importance of non-technical skills.

The evaluation results demonstrated that the WAFV, 2HAND, SJE, MMI and WCT were all suitable for inclusion in a new psychometric assessment process. VIGIL was not considered to be suitable because the results showed no evidence of criterion validity for train driving. Full descriptions of the method and results of this evaluation are provided in Annex 3.

The findings of the T948 trial were considered in conjunction with the findings from T340 and T628 in order to formulate a recommendation for the new psychometric assessment process (See Table 1 for an overview of the recommendation).

Table 1 - Overview of the recommended psychometric assessment process

Selection criterion	Assessment method
Attention	Test of Everyday Attention for Occupational assessment (TEA-Occ) Group Bourdon
Vigilance	WAFV
Memory	Trainability for Rules and Procedures 1 (TRP1)
Reasoning	Trainability for Rules and Procedures 2 (TRP2)
Perception	Adaptive Tachistoscopic Traffic Test (ATAVT)
Reaction time	WAFV
Hand coordination	2HAND
Communication	Multimodal Interview (MMI) Written Communication Test (WCT)
Conscientiousness	Situational Judgement Exercise (SJE) Multimodal Interview (MMI)
Dealing with challenging situations	
Tolerance for low stimulation	

The full recommendation with associated scoring rules is detailed in Annex 2 and the rationale for the recommendation is summarised in Section 7.

If implemented, the recommended process would provide a comprehensive assessment of the key skills against a minimum standard which would be set to exclude people who might not have the skills to be safe train drivers. The recommended scoring rules were carefully formulated with fairness in mind and it is expected that the difference in pass rates between white and ethnic minority candidates would be reduced compared to the current process.

However, many of the assessment methods recommended have not yet been used for recruitment so it is not yet known how they would perform within a candidate population. If implemented, the pass rates of the new process should be monitored and the results should be reviewed periodically to ensure that the scoring rules remain suitable from a safety and fairness perspective.

1.5 – Deliverables

This full report is the main deliverable for project T948. It contains the full details of the research method and recommendations. The project also delivered final versions of the SJE, MMI and WCT along with associated test manuals and a training package. These deliverables are only available to duty holders with responsibility for train driver assessment or recruitment.

There is also a publically available research brief and summary report that do not contain full details of the selection criteria and scoring rules because this information could give an unfair advantage to candidates.

1.6 – Next Steps

The benefits of the research can only be realised if the recommendations are implemented as a new standardised psychometric assessment process for train drivers. The next step is to progress the change through the normal standards process and plans for that activity are in hand. This includes an impact assessment to explain the benefits and costs of the change. Subject to industry approval, the recommendations would be incorporated into a new version of RIS-3751-TOM. In preparation for industry consultation on the new standard, RSSB have planned and initiated a communications programme to promote awareness of the possible changes and to allow industry to make initial preparations.

If the changes are approved, RSSB would continue to work with industry to prepare for implementation. In particular, RSSB would support the training, licensing and maintenance of the assessment methods that have been developed by RSSB. This work has already been planned and resources allocated. Rail assessment centres would also need to make the necessary preparations such as the procurement of equipment and licences, obtaining training and updating processes. Rail assessment centres are aware of this requirement via their involvement in the research steering group. An ‘implementation working group’ consisting of RSSB and assessment centre staff has been formed ready to progress the necessary actions.

2 Introduction to this report

This is the final report of RSSB R&D project T948 *Train Driver Selection Implementation Phase.* Project T948 was the final research project of a significant programme of RSSB research work in this area and has resulted in a set of final recommendations for how the standardised train driver psychometric assessment process should look in the future. This report provides a high level summary of the whole programme of research and summarises the final recommendations and rationale.

The annexes to this report provide further detailed information to support this summary. In particular, Annex 3 provides the detailed method and findings of the research that was conducted as part of project T948 specifically.

3 Background

A standardised psychometric assessment process for train drivers was introduced in 1988 by British Rail. The methods and scoring have undergone minor changes on several occasions since then but are still recognisable as the original process.

The process consists of a set of selection criteria that are considered relevant to train driving, a method for the assessment of each criterion and associated pass marks for each criterion. The purpose of the standardised assessment process is to assess whether candidates meet a safe minimum standard in attributes that are necessary for safe train driving.

The process that is current at the time of publication of this report (December 2012) is detailed in in Rail Industry Standard RIS-3751-TOM, Issue One *Train Driver Selection.* Rail Industry Standards are not mandatory. They are produced by RSSB at the request of industry where there are expected to be benefits from different companies using a common standard. In the case of the driver psychometric assessment process, the standard is widely adopted. Train driver assessments are conducted by a small number of company assessment centres and private service providers. Assessment centres that are following the process in RIS-3751-TOM, Issue One are members of the Rail Assessment Centre Forum (RACF).

There are several advantages in having a standard psychometric assessment process that is consistently applied during initial selection regardless of employer. It is an efficient way to measure underlying abilities that cannot be observed during traditional interviews. This contributes to a safe, consistent minimum standard of ability in all drivers. No other competence or medical assessment is likely to detect underlying deficiencies in some of the attributes (such as attention) once a driver is qualified.

All potential drivers are assessed against the same standard. Therefore, drivers can easily transfer between companies without needing to be assessed again. Having a standardised process means companies do not have to design their own process individually.

There are several reasons why the standard process needs to be updated now. Firstly, the selection criteria need to be updated. More detail about this is provided in Section 5. Secondly, if the selection criteria are updated, including the addition of new criteria, then new assessment methods are needed to cover all the criteria. Finally, there are some problems with the current assessment methods. Project T340 (RSSB, 2005) evaluated the current assessment methods and recommended that new assessment methods needed to be identified due to inconsistent evidence of validity of the current assessment methods and to incorporate current good practice (see Section 4 for more details).

There is also a need to ensure that the process complies with the Equality Act 2010 (HMSO, 2010). The fairness of the psychometric assessment process has been challenged in the past. In 1990 a group of Paddington guards alleged that the psychometric tests discriminated against ethnic minorities and the case was settled out of court. As part of the settlement, British Rail agreed to work with the Commission for Racial Equality (CRE) to develop a fairer assessment procedure, monitor its recruitment of train drivers and consider positive action for ethnic minorities. British Rail worked with the CRE for two years to monitor success rates of minority candidates and implemented several programmes of 'access training' designed to assist candidates to prepare for the train driver selection process. A report on this work was published in 1996 (CRE, 1996).

In 2011, RSSB examined the pass rates for each psychometric assessment method and found that significantly lower percentages of black and Asian candidates pass the assessment process compared to white candidates and that it could be considered to be unfair (RSSB, 2011a).

The industry, via the Driver Selection Governance Group, are committed to addressing these limitations within the constraint of needing to have a psychometric assessment process that effectively filters out candidates who do not have the aptitude to become safe train drivers. Industry strategy regarding train driver psychometric assessment is published on the RSSB website (RSSB, 2011b).

The overall objectives of the industry's work on train driver selection are:

- To implement an updated process for train driver psychometric assessment in RIS-3751-TOM
- To monitor the effectiveness of this process on an on-going basis so that it can be maintained fit-for-purpose.

RSSB has a remit to assist industry to achieve these aims by making recommendations about what the updated process should be and by undertaking reviews of the process to inform industry monitoring.

4 The research work

RSSB's work in this area has primarily been addressed via the Research and Development Programme and there have been three key projects. This section of the report outlines the key findings and recommendations of the three projects.

Project T340 *Psychometric testing – A review of the train driver selection process* (RSSB, 2005) was conducted by CAS (Competence Assurance Solutions) and was published in 2006. T340 had a wide scope that covered the whole driver selection process, from application to final recruitment decision, and other uses of psychometric assessment such as post-incident. Only the aspects of T340 relating to psychometric assessment during recruitment are relevant and will be discussed here.

Research project T340 identified some areas of weakness in the current psychometric assessment process. In particular, validity results for the DTG and Group Bourdon were inconclusive. There were some correlations with job performance measures but not always where expected and relationships were not consistently demonstrated in different studies that were reviewed. CAS concluded that the inconsistency in the validity findings over time for the Group Bourdon and DTG suggested that these parts of the process do have some value but the predictive validities were probably low. This would explain why relationships were not found in every validation study. The review also showed that the difference between score bands being used to make selection decisions was not always statistically significant. Different versions of assessment methods were not equivalent. The review found there was not much evidence of validity for most sections of the Criterion Based Interview (CBI). At that time, the CBI had a very high pass rate so there was also a question over whether the method was adding enough value to justify the time spent on it.

T340 also found that the selection criteria used were in need of updating and identified an initial set of new selection criteria that provided a better coverage of the key requirements.

The project made the following relevant recommendations:

- The selection criteria should be updated to give better coverage of the abilities recognised to underpin good performance in modern train driving.
- Selection criteria which address safety and train handling performance should form the core of driver selection and be assessed by all companies in the same way either by the

assessment centres or by qualified individuals in companies.

- Companies should have flexibility in the way in which they assess criteria which relate to personal effectiveness.
- The effective parts of the current assessment centre process should be retained for national use.
- The parts of the assessment centre process which have not proved effective should either be replaced or upgraded.
- New tests should be identified or developed to assess the new or extended selection criteria that have been identified.

In response to the recommendations of T340, RSSB undertook research project T628 *Train Driver Selection – Development Phase* (RSSB, 2010). The objectives that are relevant to updating the train driver psychometric assessment process were:

- Further develop the train driver selection criteria proposed in T340 and obtain industry agreement for the use of these selection criteria.
- Propose a revised train driver psychometric assessment process which will effectively measure the updated selection criteria, conform to good practice in selection, exclude unsuitable drivers and where possible enable potentially exceptional drivers to be distinguished.
- Evaluate the arrangements for the management of the train driver psychometric assessment process and propose alternative arrangements if weaknesses are apparent.

As part of T628, RSSB further developed the selection criteria that were initially proposed in project T340 (see Section 5). New assessment methods were shortlisted as potential tests to replace the DTG and Group Bourdon and to assess the new selection criteria that had been added. RSSB conducted a trial of shortlisted assessment methods to evaluate their reliability, validity and other attributes (see Table 2 for a list of assessment methods evaluated in the trial). As part of T628, RSSB also developed a bespoke Situational Judgement Exercise (SJE) and Multi-Modal Interview (MMI) to measure the specific behavioural attributes described in the new selection criteria.

Table 2 – Assessment methods trialled during project T628

Selection criteria	Assessment methods trialled	Publisher
Attention	The Test of Everyday Attention for Occupational Assessment (TEA-Occ) The Simultaneous Capacity and Stress Tolerance Test (SIMKAP)	Pearson Assessments Schuhfried
Vigilance	TEA-Occ	Pearson Assessments
Reasoning	SIMKAP	Schuhfried
Perception	The Tachistoscopic Traffic Test (TAVTMB) The Time Movement Anticipation Test (ZBA)	Schuhfried

Research project T628 concluded that all of the assessment methods trialled had the potential to be implemented and provided information regarding the reliability, validity, fairness, administration and cost of the methods. Based on the trial results, the TEA-Occ was not considered to be suitable for the assessment of vigilance, so it was recommended that a measure of vigilance be identified and evaluated in the future. During the course of the T628 project, hand coordination was added to the selection criteria in response to the requirements of the European Commission Directive 2007/59/EC, now transposed into national law as the Train Driver Licence and Certificates Regulations 2010 (TDLCR, 2010). A measure of hand coordination had not been included in the T628 trial so it was recommended to identify and evaluate a test for this purpose in future work. Finally, during T628 the steering group for the project requested RSSB to develop a means of assessing written communication that would be more structured than the current means. Written communication is currently assessed using the pre-interview form for the CBI.

To address the final objective of project T628 a review of the management arrangements for train driver psychometric assessment was commissioned. The review concluded that there was no strategic oversight of the process and no clear mechanism for driving forward improvements (Arthur D Little, 2009). It recommended that a new industry committee be established to monitor the effectiveness of the train driver psychometric assessment process and develop an industry strategy. This recommendation was taken forward in parallel to the research work.

Following on from project T628, there was a need to identify methods to assess the outstanding selection criteria (vigilance and hand-coordination) and to evaluate the bespoke behavioural assessment methods that had been developed, including the new Written Communications Test (WCT). Therefore, project T948 *Train Driver Selection – Implementation Phase* was initiated to address these outstanding tasks. The objectives of project T948 were to:

- Evaluate the suitability of the proposed methods for measurement of behavioural criteria, communication, hand coordination and vigilance, in terms of validity, reliability, fairness and practicality.
- Propose a revised assessment centre process that would effectively measure the updated selection criteria, conform to good practice in selection, exclude unsuitable drivers and where possible enable potentially exceptional drivers to be identified.

Research project T948 was conducted by RSSB. It included another evaluation trial of potential new assessment methods (see Table 3 for a list of the assessment methods trialled). Section 1 of Annex 3 provides a full description of each of the assessment methods trialled.

Table 3 – Assessment methods trialled during project T948

Selection criteria	Assessment methods trialled	Publisher
Vigilance	WAFV VIGIL	Schuhfried
Hand coordination	2HAND	Schuhfried
All behavioural selection criteria	Situational Judgement Exercise (SJE) Multimodal Interview (MMI)	RSSB
Written communication	Written Communication Test (WCT)	RSSB

The final stage of research project T948 was to propose a revised assessment centre process to measure all the updated selection criteria, taking into account the findings from T340, T628, T948 and any other evidence gathered by the project team during the course of the research.

The outcomes of this programme of research in terms of the final proposed selection criteria, assessment methods and scoring rules are outlined in the following sections of this report. Further detailed information that underpins the recommendations and rationale is provided in the annexes to this report and the published reports for T340 and T628.

5 The new selection criteria

The current selection criteria that are detailed in RIS-3751-TOM *Train Driver Selection, Issue One,* are a mix of safety, performance and trainability criteria pitched at varying levels of generality. They have remained largely unchanged since psychometric assessment was first introduced in 1988. They require updating in light of the evolution of the train driver role over time and to integrate requirements of the TDLCR (2010).

The new recommended selection criteria were developed over the course of the three research projects in conjunction with the industry steering group and in response to emerging knowledge from the evaluation trials and the finalisation of the TDLCR (2010).

The recommended selection criteria were developed to:

- Match the train driving role – Involvement of subject matter experts was used to design a set of criteria that cover the key skills required for train driving now. The primary changes were the inclusion of vigilance as a criterion and greater emphasis on behavioural attributes.
- Reflect the importance on non-technical skills – A growing body of research demonstrates how behaviours underpin technical performance (Flin, O'Connor, & Crichton, 2008; RSSB, 2012). There is strong interest in this within the GB rail industry. Better assessment of these skills at the recruitment stage is part of an overall ambition to focus on non-technical skills throughout the workforce development cycle.
- Satisfy legislation – TDLCR (2010) requires certain criteria to be assessed prior to employment. These criteria were incorporated and the language used to describe them was designed to provide transparent compliance.
- Be consistent with psychological theory – A wealth of psychological literature describes cognitive constructs such as attention. Where possible, the design and description of the selection criteria took into account the relevant theories.
- Have a consistent level of description and definition – Each of the proposed selection criteria is described at a consistent level of generality and is accompanied by a clear and precise definition that would assist duty holders and candidates to understand what is being assessed.

Table 4 outlines the new selection criteria and shows how they relate to the existing selection criteria and the selection criteria that are required to be assessed by TDLCR (2010). A full definition of each criterion is given in Table 16.

Table 4 - Recommended selection criteria in comparison to requirements of TDLCR 2010 and the existing selection criteria

Category	Requirements of the TDLCR (2010)	Existing selection criteria in RIS-3751-TOM	Recommended new selection criteria
Cognitive	Attention	Ability to maintain vigilance and concentration at all times	Attention
	Concentration	Ability to maintain vigilance and concentration at all times	Vigilance
	Memory	Ability to retain and recall job related information Ability to learn new information	Memory
	Perception	-	Perception
	Reasoning	-	Reasoning
	Communication	Ability to communicate clearly and effectively	Communication (including verbal communication and written communication)
Psychomotor	Reaction time	Ability to react safely and quickly	Reaction time
	Hand coordination	Ability to operate a range of hand and foot controls	Hand coordination

Category	Requirements of the TDLCR (2010)	Existing selection criteria in RIS-3751-TOM	Recommended new selection criteria
Behavioural	No established occupational psychological deficiencies, particularly in operational aptitudes or any relevant personality factor	Motivation to follow set rules and procedures Conscientiously works to exceed training course demands Checks, does not make assumptions	Conscientiousness (incorporating sub-criteria: dependability; attitude to work and people; commitment to work; attention to detail; ability to check and not make assumptions; compliance with rules and procedures)
		Ability to remain calm in emergency/stressful situations Is proactive and tenacious	Dealing with challenging situations (incorporating sub-criteria: proactivity; tenacity; assertiveness; calmness under pressure; reactivity to stress)
		Work alone for long periods, maintaining the appropriate standards of competence at all times	Tolerance for low stimulation (incorporating sub-criteria: social need; sensation seeking; need for external stimulation)

6 Principles of building a psychometric assessment process for selection

The overall objective of the project was to recommend an updated psychometric assessment process that would provide a reliable, valid, fair and practicable assessment of the aptitudes required for train driving against a defined minimum standard. The recommended psychometric assessment should result in an overall pass or fail decision for each candidate. That decision would be based on scoring criteria that are agreed by industry and applied consistently by all duty holders who use the process. The development of this core assessment was the main focus of the research programme. The project sought to define three aspects of the selection process:

1. The assessment methods to use for the assessment of each selection criterion.

2. The particular scores to use for the assessment of each selection criterion.
3. How to make pass and fail decisions for each selection criterion including the particular cut-offs to use for individual scores.

Candidate train drivers need to demonstrate that they have attained a minimum standard on every safety and time critical ability and behaviour. In other words, the assessment of safety and time critical abilities should result in a pass or fail decision. Safety and time critical abilities and behaviours are those that need to be demonstrated at all times while driving a train. Failure to demonstrate these skills within an appropriate time limit while driving a train could result in an error with safety consequences, such as a signal passed at danger (SPAD). All of the selection criteria are considered to be safety and time critical except for written communication.

Written communication skill is important to the train driver role and relates to safety but most driver written tasks are completed at the depot or during breaks. Therefore, they are not time critical and lack of skill cannot result in a train driving error.

A recommended psychometric assessment process and pass/fail criteria were designed taking into account all the available evidence including the results of validation trials from T340 (RSSB, 2005), T628 (RSSB, 2010) and T948 and other evaluations of potential psychometric assessment methods that were available to RSSB in sufficient detail. In order for an assessment method to be used fairly to make pass and fail decisions it is important that it demonstrates sufficient levels of reliability and validity.

The assessment should also provide qualitative information about the strengths and weaknesses of each candidate that could be used as required by each company. This aspect is additional to the core safety requirement for psychometric testing but has been requested by duty holders so that they can distinguish between candidates who pass and potentially use the information for on-going driver management. The details of how the assessment results can be used in this way are not discussed in detail in this report and would not be part of the core process defined in RIS-3751-TOM because each duty holder would choose if and how they use this additional information.

6.1 Considerations for recommending assessment methods

The decision about which assessment methods to recommend for measurement of each selection criterion took into account the following considerations:

- Coverage of the selection criteria – Each of the individual selection criteria should be assessed. Where a selection criterion is made up of distinct sub-criteria, each sub-criterion should be assessed.

- Use of multiple independent measures – Where possible, each selection criterion should be assessed by more than one assessment method to provide increased reliability of measurement.

These considerations are recognised as good practice in the design of assessment centres (see, for example, the BPS, 2006).

6.2 Considerations for recommending particular scores

In practice the recommendation of what assessment methods to use was led by consideration of what particular scores to use, taking into account the following considerations:

- Reliable and valid psychometric assessment scores – The evaluation evidence for each assessment method score should demonstrate acceptable levels of reliability and validity.
- Each psychometric assessment score provides unique information – The process should not include measures that are so highly correlated with other measures that they do not provide any added value to the assessment.

6.3 Considerations for specifying pass/fail decisions

The recommendations for how to reach the pass/fail decision for each selection criterion took into account the following considerations:

- Pass marks screen out potentially unsafe candidates – The pass marks should be designed to assess candidates against a minimum standard where scores below the pass mark indicate that the candidate might not have sufficient aptitude to perform safely. They should not be set to identify potentially exceptional candidates.
- A reasonable overall pass rate for each assessment method – The number of candidates screened out by each assessment method should be reasonable given the accuracy with which each assessment method can distinguish between good and poor candidates.

6.4 Considerations regarding the design of the overall process

Safety was the primary consideration when formulating the scoring rules and took priority over all other considerations. Within this constraint, the recommended overall assessment process was designed to have the following attributes as far as possible:

- Fairness – The predicted pass rates for minority ethnic groups, females and older candidates should not be less than 4/5^{ths} of the predicted pass rate of the majority groups.
- A reasonable overall pass rate for the process – Practicalities of recruitment are such that a certain percentage of candidates need to pass in order to have

enough candidates to choose from for the roles available. Too many or too few candidates passing the process would cause problems.

- Ability to sift candidates based on paper tests – Assessment centres have a limited number of computers. Some assessment centres need to assess large numbers of candidates in a short space of time. Therefore, they need to be able to screen out candidates based on assessment methods that can be administered to a large group.
- Assessment can be conducted in one day – The overall time taken to go through the whole assessment centre process should be less than one working day.

The recommended psychometric assessment process was designed to strike a suitable balance between all of these requirements and is summarised in Section 7.

7 The recommended psychometric assessment process

The recommended psychometric assessment process is detailed in Annex 2. The following sections summarise the rationale for the design of the assessment process for each selection criterion. Each section starts with a statement about what has been recommended and a summary rationale follows. The full detailed rationale including statistical figures relating to the evaluation of each psychometric assessment method and the approach to scoring decisions is included in Annex 3 Section 7.

The recommended psychometric assessment is presented below with a section for each selection criterion. The order of the sections represents a suggested order for administering the assessment process. The design of the order took into account the practical requirement to be able to sift large groups of candidates using paper tests and the order of importance of the selection criteria. The vigilance assessment is recommended to be the first of the computerised assessments because it should be conducted before candidates become too fatigued. The behavioural assessment methods are the most time consuming so they should be conducted last.

The predicted pass rates for each selection criterion were roughly estimated using norm data and data regarding the correlations between assessment method scores where available. The predicted pass rates were based on the assumption that the assessment would be conducted in the order specified and candidates who fail on a selection criterion would be screened out of subsequent assessments. Norm and correlation data was not available for all assessment scores and the trial data was from a pool mostly consisting of train drivers not real candidates. Therefore, the predicted pass rates should be treated as an educated guess.

7.1 Attention

7.1.1 Definition

Selective attention - The ability to differentiate between different sources of information and attend selectively to them, eg distinguishing and attending to alarms.

Divided attention - The ability to switch attention between sources of information, eg lineside information and in-cab information and perform different tasks in parallel, such as making train announcements while on the move.

7.1.2 Recommendation

The paper Group Bourdon and the TEA-Occ lift counting with distraction and telephone search with counting sub tests are recommended for the assessment of attention (See Table 5).

Table 5 – Attention criterion – Recommended assessment scores

Selection criterion	Assessment method	Score	Raw score pass criteria
Attention	TEA-Occ	Lift counting with distraction – The number of correctly counted strings of tones.	≥ 6
		Dual task decrement – The extent to which performance is reduced when doing two tasks simultaneously.	≤ 4.44
	Paper Group Bourdon	Total production – The number of stimuli that are checked within the time limit.	≥ 938
		Total omissions – The number of target stimuli that are missed.	≤ 47

7.1.3 Coverage of the selection criteria

There were three potential assessment methods for measuring attention: Group Bourdon, TEA-Occ and SIMKAP. The evidence of criterion validity for SIMKAP was weak and fairly inconsistent. Therefore SIMKAP was discarded from consideration.

The Group Bourdon could not be used alone because whilst it provides a measure of visual selective attention, it does not assess divided attention and this is an important ability for train drivers. The TEA-Occ assesses both auditory selective attention and divided attention so is recommended for use in conjunction with the Group Bourdon.

7.1.4 Use of multiple independent measures

Evidence in favour of the TEA-Occ was not sufficiently strong to justify replacing the Group Bourdon completely because some subtests had stronger criterion validity than others. It is therefore proposed that a customised (and shortened) version of the TEA-Occ is used. The lift counting with distraction and telephone search while counting subtests displayed evidence of construct and criterion validity and are recommended to supplement the Group Bourdon. The approach of using the Group Bourdon and the TEA-Occ in combination would provide two points of measurement for attention and would cover both of the attention sub-criteria. It would therefore be a more robust solution than using one test alone.

7.1.5 Reliable and valid psychometric assessment scores

Based on the results of the T628 validation study (RSSB, 2010), the main scores from the TEA-Occ lift counting with distraction and telephone search with counting sub-tests were considered suitable for use as part of the process. However, the evidence from this single study only showed weak criterion validity, so it would be appropriate to use the TEA-Occ in combination with the Group Bourdon for this important selection criterion.

Evaluation evidence for the Group Bourdon was inconsistent with some studies showing relationships to relevant job performance measures and some not. The errors score is currently in use but there was little evidence of its criterion validity. Therefore, it is recommended to make use of only the productions and omissions scores in future.

7.1.6 Each psychometric assessment score provides unique information

Each score recommended for use is derived from a different task and was designed to measure a slightly different aspect of attention. The correlations between the test scores were low enough to suggest it would be worthwhile including all of them in making the pass/fail decision because they would all provide unique information within the psychometric assessment process.

7.1.7 Pass marks screen out potentially unsafe candidates

For the attention criterion the objective is to screen out candidates who have a very poor ability to maintain attention. Each individual score measures a slightly different aspect of attention. Therefore, it is recommended to use conjunctive scoring rules where candidates need to reach a minimum level on each individual score. The recommended minimum level for each individual score was set relatively low because attention would be assessed by four different scores.

The current pass marks for the Group Bourdon were used as a starting point for recommending pass marks for the updated

process. It was found that the omissions pass mark was set very high and excluded many more candidates than could be justified on the basis of the validity of this particular score. In addition, there was a strong difference in performance on this score between white and ethnic minority candidates which resulted in a much lower pass rate for minority candidates. Therefore, given that the Group Bourdon would be supplemented by another test of attention, it is recommended to significantly reduce the omissions score pass mark to a level that would screen out approximately the bottom 10% of candidates. The production score pass mark is recommended to remain as it is currently and would screen out approximately the bottom 11% of candidates.

The recommended pass marks for Group Bourdon were applied to one year's previous assessment centre data to assess the effect on the pass rate. The difference in pass rates between different ethnic groups was significantly reduced. The safety disbenefit that could arise from reducing the pass mark would be mitigated by the inclusion of the TEA-Occ as an additional measure of attention.

The pass mark for the TEA-Occ lift counting with distraction score was determined based on information from the test publisher relating to the average performance of normal and brain injured patients on the TEA. The recommended pass mark was set at a level that should identify people with poor attention who score much lower than the average for people without brain injury. The available information for the dual task decrement score was not so clear. The largest norm group available was a group of signallers and train drivers (n=132) who were assumed to have better than average attention because they were already assessed and trained to maintain attention. The recommended pass mark for dual task decrement was set at a level that is better than the lowest 5% of the signallers and drivers assessed. Based on a smaller sample from the general population, it is estimated that this pass mark would exclude approximately 10% of candidates.

7.1.8 Reasonable overall pass rate for each assessment method

The paper Group Bourdon pass marks would screen out a maximum of 21% of candidates. The actual figure would probably be lower because there is a relationship between the production and omissions scores which means that people with acceptable levels of attention would be likely to reach the minimum standard on both scores.

There were no role specific norm tables available for the TEA-Occ (though general population norms were available for the TEA) so it was not possible to accurately estimate what percentage of candidates would be screened out using the recommended cut-offs. However, given that they were set conservatively to identify people with very poor attention, the percentage should not be too high.

The research estimated that a maximum of 35% of candidates would be screened out on the basis of the recommended attention assessment using a combination of the Group Bourdon and TEA-Occ. This would be a similar level to the current process.

7.2 Memory

7.2.1 Definition

The ability to learn, recall and apply job related information in appropriate time limits, eg learn new information in training; remembering instructions from signallers; applying specific rules and procedures.

7.2.2 Recommendation

The TRP1 assessment method with a pass mark of nine or above is recommended to assess memory. The score is the number of correctly answered questions.

7.2.3 Coverage of the selection criteria

The TRP1 assessment method measures the ability to learn fictitious rules information, remember it and apply it in order to answer a series of questions. The TRP1 test is in use already as part of the current assessment process.

No other measures of memory were considered because the validation evidence from previous research was strong for TRP, it was designed specifically for use as part of train driver psychometric assessment and there were no practical reasons not to continue to use it.

7.2.4 Use of multiple independent measures

No other methods in the set evaluated were suitable to assess memory specifically. However, several of the other recommended assessment methods have a memory element.

7.2.5 Reliable and valid psychometric assessment scores

The TRP1 displayed good and consistent evidence of criterion validity for train driving. Scores on the TRP1 were particularly related to driver training outcomes.

7.2.6 Each psychometric assessment score provides unique information

Only one score is recommended to assess memory. The TRP1 score was not highly correlated with any other assessment method score that is recommended for use because it was considered to provide unique information within the psychometric assessment process.

7.2.7 Pass marks screen out potentially unsafe candidates

The current pass mark for TRP1 was used as a starting point. Using the current pass mark, the pass rate for black candidates was statistically lower than the pass rate for white candidates and could

be considered to be unfair, especially since the results of this assessment method primarily related to training outcomes and not to safety performance. Therefore, it is recommended to relax the pass mark by one point. When applied to previous assessment centre data this change reduced the difference in pass rates but the pass rate for ethnic minority candidates was still lower at 85% compared to 97% for white candidates.

7.2.8 Reasonable overall pass rate for each assessment method

Based on previous assessment centre data, it was estimated that a maximum of 6% of candidates would fail the memory assessment. However, TRP1 would be performed following the attention assessment and it was predicted that up to 4% of candidates who passed attention would fail on memory.

7.3 Reasoning

7.3.1 Definition

The ability to solve problems and make decisions, eg fault diagnosis; understanding and interpreting information from instrumentation.

7.3.2 Recommendation

The TRP2 assessment method with a pass mark of 13 or above is recommended for the assessment of reasoning. The score is the number of correctly answered questions.

7.3.3 Coverage of the selection criteria

There were two potential assessment methods for measuring reasoning: TRP2 and SIMKAP. The evidence of criterion validity for SIMKAP was inconsistent. Therefore SIMKAP was discarded from consideration.

The TRP2 is a suitable assessment of reasoning because it requires the candidate to apply information they have learnt in order to make decisions. The TRP2 is in use already as part of the current assessment process.

7.3.4 Use of multiple independent measures

No other methods in the set evaluated were suitable to assess reasoning specifically.

7.3.5 Reliable and valid psychometric assessment scores

Evidence from various sources consistently demonstrated a relationship between performance on the TRP2 and training and safety performance.

7.3.6 Each psychometric assessment score provides unique information

Only the TRP2 score would be used to assess reasoning. The TRP2 was moderately correlated with the TRP1 but not so highly that no new information would be provided by the TRP2.

7.3.7 Pass marks screen out potentially unsafe candidates

Again, the pass mark currently used for the TRP2 was used as a starting point in setting the pass mark for use within the new recommended process. Pass rates using the current pass mark were significantly lower for black candidates than for white candidates. It is recommended to relax the pass mark by one point to reduce this difference. This recommended cut-off was applied to previous assessment centre data and pass rates were in compliance with the four-fifths rule. However, it is likely that there would still be a difference in pass rates between blacks and whites. Given the strong relationship between results on this assessment method and performance on the job, it was not considered justifiable to relax the pass mark further due to safety concerns.

7.3.8 Reasonable overall pass rate for each assessment method

The new recommended pass mark was set at a point that would screen out approximately 11% of applicants if administered alone. However, the TRP2 would be administered after the attention and memory assessments so the actual failure rate for reasoning was predicted to be around 3-4%.

7.4 Vigilance

7.4.1 Definition

The ability to attend and respond to stimuli which occur relatively infrequently and over extended periods of time.

7.4.2 Recommendation

The WAFV is recommended to assess vigilance. Two different scores are recommended as shown in Table 6.

Table 6 – Vigilance criteria – Recommended assessment scores

Selection criterion	Assessment method	Score	Raw score pass criteria
Vigilance	WAFV	Missed reactions – The number of target stimuli that are missed after 1500ms of presentation.	≤ 5
		False alarms – The number of non-target stimuli that are incorrectly responded to.	≤ 8

7.4.3 Coverage of the selection criteria

The WAFV was specifically designed to measure the ability to detect events that occur infrequently.

7.4.4 Use of multiple independent measures

No other measures in the set evaluated were suitable to assess vigilance specifically.

7.4.5 Reliable and valid psychometric assessment scores

The results of the T948 trial showed strong evidence of criterion validity. The missed reactions and false alarms scores were significantly related to relevant measures of safety performance on the job.

7.4.6 Each psychometric assessment score provides unique information

The missed reactions score was recommended by the test publisher as the primary measure of vigilance. The missed reactions and false alarms scores were moderately correlated but both were necessary to have a 'cheat-proof' assessment of vigilance. The false alarms score was required because otherwise a candidate could pass by responding to *every* stimulus whether target or non-target and this would not be a valid measure of vigilance.

7.4.7 Pass marks screen out potentially unsafe candidates

Vigilance is not currently specifically identified as a selection criteria for train drivers. It was added as a specific selection criterion because it was considered to be one of the most critical abilities for safety performance. The proposed pass mark was determined based on the poorest levels of performance observed in the trial sample. At the level proposed, trial participants who would fail if it was applied to them tended to be those with errors and / or operational incidents in their records. The pass mark for missed reactions was proposed at a level that should exclude the worst 16% of performers based on the norm group supplied by the test publisher. It was set quite high due to the importance of this selection criterion. The false alarms score was set to exclude the worst 5% of performers because the intention was to exclude those who wilfully attempt to cheat the test by pressing the button repeatedly. This would also exclude those who genuinely cannot distinguish between target and non-target stimuli.

7.4.8 Reasonable overall pass rate for each assessment method

The vigilance assessment would be administered following the attention, reasoning and memory assessments. It was predicted that 9-10% of the remaining candidates would fail the vigilance criterion with the recommended scoring rules.

7.5 Reaction time

7.5.1 Definition

A quick and adequate response to simple and complex visual and acoustic stimuli and the associated quality of performance.

7.5.2 Recommendation

The reaction time score from the WAFV with a pass mark of 656 milliseconds or faster is recommended for the assessment of reaction time.

7.5.3 Coverage of the selection criteria

The WAFV reaction time score is recommended for the assessment of reaction time because it provides a measure of simple reaction time where the candidate responds to a simple stimulus as quickly as possible.

The reaction time selection criterion includes the ability to react to both simple and more complex stimuli and to give an appropriate response. The working time score of the TAVTMB that was trialled during project T628 was also considered and would have been recommended as an assessment of reaction to more complex stimuli. Unfortunately, the TAVTMB was updated by the test publisher and the new version, the ATAVT is an adaptive test. In the ATAVT, the working time measure is not standardised for all participants because the assessment takes a different amount of time depending on how the candidate performs. Therefore, it could not be used for the assessment of reaction time and the WAFV needed to be used alone.

7.5.4 Use of multiple independent measures

As explained above, it was not possible to use multiple independent measures for the reaction time criterion.

7.5.5 Reliable and valid psychometric assessment scores

Analysis from the T948 validation trial showed that the WAFV reaction time score was significantly related to safety performance.

7.5.6 Each psychometric assessment score provides unique information

The WAFV mean reaction time score was significantly correlated with the other WAFV scores as would be expected. It was significantly but not strongly correlated with two other scores that were proposed for use. It is the only measure of those evaluated that directly relates to speed of reaction so it would provide unique information.

7.5.7 Pass marks screen out potentially unsafe candidates

The pass marks for reaction time were set taking into account the requirements of the train driving task. Train drivers need to react within a reasonable amount of time but they do not need to have exceptionally fast reactions. To provide an example, train drivers have either 2 or 2.7 seconds to respond to AWS depending on whether the train is a high speed one or not. Therefore, the pass mark for WAFV reaction time was set to exclude only the worst 5% of performers.

7.5.8 Reasonable overall pass rate for each assessment method

The pass mark for the reaction time score was set quite low. It was estimated that the reaction time assessment would exclude approximately 4-5% of candidates who have not been excluded by the previous assessments.

7.6 Perception

7.6.1 Definition

The ability to anticipate elements in a traffic environment and make a correct decision about how to respond given the speed and distances involved, eg identifying a landmark cue before a station and starting to decelerate.

7.6.2 Recommendation

Perception is recommended to be assessed using the overview score from the ATAVT with a pass mark of greater than -1.5066. The ATAVT is the updated version of the TAVTMB that was trialled during the T628 project. The original TAVTMB is no longer available but information in the ATAVT test manual provided by the test publisher shows that the overview scores are equivalent between the two tests. The evaluation findings therefore relate to the TAVTMB but the recommendation specifies the ATAVT.

7.6.3 Coverage of the selection criteria

The ATAVT provides a suitable measure of the speed of perception and the ability to perceive elements within a traffic environment. As it is described in the test manual as a measure of speed of perception, it also assesses the ability to see and recognise objects in the environment after short glimpses.

The ZBA time movement anticipation test was trialled as a potential measure for this latter component of the perception criterion during the T628 project. However, the results suggested that the ZBA had poor face and criterion validity for train driving so it is not recommended for use.

A specific assessment of time movement anticipation is not included in the current assessment process. It is recommended to take the assessment of perception forward using just the ATAVT and if it emerges that there is a significant gap in trainee driver skills in this area then an additional assessment method can be trialled and introduced at a later date.

7.6.4 Use of multiple independent measures

The ATAVT is the only specific measure of perception recommended for inclusion in the new train driver psychometric assessment process. However, several of the other assessment methods require perceptual skills to perform well on them even if this is not the specific purpose of the assessment.

7.6.5 Reliable and valid psychometric assessment scores

Results from the T628 evaluation and analysis during T948 using updated job performance data consistently showed good relationships between the TAVTMB overview score and various job performance measures. The test publishers stated that results of validation studies can be directly generalised from the TAVTMB to the ATAVT.

7.6.6 Each psychometric assessment score provides unique information

The ATAVT overview score is the only specific measure of perception in the recommended assessment process and is clearly different from other scores that are recommended. The equivalent TAVTMB score was correlated with some of the other scores that are recommended for use but not to an extent that would make it redundant.

7.6.7 Pass marks screen out potentially unsafe candidates

The TAVTMB overview score showed a strong and consistent relationship with job performance so it was justifiable to set the pass mark at a higher level than for some of the test scores where the evidence of criterion validity was weaker. The recommended pass mark was set to exclude the worst 13% of performers on this test based on the norm group provided by the test publisher. The train drivers in the trial sample were considered to be adequate performers who were already assessed for perception, albeit using a different method. The recommended pass mark equated to the lowest scores achieved by the train drivers in the trial sample.

7.6.8 Reasonable overall pass rate for each assessment method

Although the percentile rank of the recommended pass mark is quite high, candidates will have already been assessed on most of the other criteria so the perception assessment should not exclude too many candidates. It was estimated that the perception assessment would exclude approximately 7% of candidates who were not excluded by the previous assessments.

7.7 Hand coordination

7.7.1 Definition

The ability to make appropriate and controlled movements in response to decisions about complex stimuli.

7.7.2 Recommendation

Hand coordination is recommended to be assessed using the 2HAND overall mean duration and percent error duration scores as shown in Table 7.

Table 7 – Hand coordination criteria – Recommended assessment method

Selection criterion	Assessment method	Score	Raw score pass criteria
Hand coordination	2HAND	Overall mean duration – The average time taken to complete the tracks.	< 52.8
		Percent error duration – The percentage of time spent outside the track.	< 16.7

7.7.3 Coverage of the selection criteria

The content of the assessment method is very focused on hand coordination and covers it well.

7.7.4 Use of multiple independent measures

The 2HAND is the only measure of hand coordination recommended for inclusion in the new train driver psychometric assessment process.

7.7.5 Reliable and valid psychometric assessment scores

The overall mean duration score demonstrated good relationships with measures of job performance that relate to train handling. The percent error duration score did not show evidence of criterion validity but needed to be included to guard against candidates cheating the assessment.

7.7.6 Each psychometric assessment score provides unique information

The 2HAND assessment is the only assessment of motor skills included in the recommended assessment process. Overall mean duration was not significantly correlated with any other recommended assessment score except for the 2HAND percent error duration.

Percent error duration was required in addition to overall mean duration otherwise a candidate could pass the assessment by quickly moving the ball to the other end of the track without attempting to stay within the lines.

7.7.7 Pass marks screen out potentially unsafe candidates

The hand coordination selection criterion was imposed by the TDLCR (2010) and was not identified by GB subject matter experts as a key safety critical aptitude for train driving. Therefore the pass mark was set low to exclude only the worst 5% of candidates on both scores.

7.7.8 Reasonable overall pass rate for each assessment method

The 2HAND is recommended to be the final part of the cognitive and psychomotor assessment. It was predicted that it would exclude up to 4-5% of the remaining candidates.

7.8 Conscientiousness

7.8.1 Definition

Conscientiousness – Has the drive and willingness to achieve the goals they are set and to complete work to the highest possible standard, working effectively with others as required.

Conscientiousness is broken down into the following dimensions:

Dependability - Can be relied upon to carry out the tasks and fulfil the responsibilities expected of them.

Attitude to work and people - Considerate, supportive and co-operative towards others.

Commitment to work - Does a task well and to best of ability, prioritising work over other commitments as appropriate.

Attention to detail - Is thorough in accomplishing a task and pays close attention to detail. Takes a systematic and unhurried approach.

Ability to check and not make assumptions - Checks own understanding of all relevant information, and others' understanding as appropriate.

Compliance with rules and procedures - Follows rules and procedures, understands their relevance and takes action if others do not follow rules or if rules are inappropriate.

7.8.2 Recommendation

The conscientiousness criterion is recommended to be assessed using the Situational Judgement Exercise (SJE) conscientiousness score and the Multi-modal interview (MMI) conscientiousness score as shown in Table 8.

Table 8 – Conscientiousness selection criterion – Recommended assessment methods

Selection criterion	Assessment method	Score	Pass criteria
Conscientiousness	SJE	Conscientiousness main criterion average score	There is no minimum requirement but the score informs the MMI. A z100 score[1] of < 77.5 = 'low' banding which requires a higher MMI mark to pass the Conscientiousness criteria.
	MMI	Topic area 1 score	≥ 3
		Topic area 2 score	≥ 3
		Topic area 3 score	≥ 3
		And if 'low' band on SJE:	
		Main criterion average	≥ 4

7.8.3 Coverage of the selection criteria

The SJE and MMI were both specifically designed to assess conscientiousness according to the agreed definition and sub-criteria so the entire criterion would be well covered.

7.8.4 Use of multiple independent measures

The SJE and MMI would provide two independent measures of the conscientiousness criterion. The SJE would be used to inform the MMI. Candidates who score in the 'low' or 'moderate' band on the SJE would be subject to situational questions in the interview as well as behavioural questions. Behavioural questions relate to past experiences and situational questions relate to hypothetical scenarios. Further information regarding the MMI is available in Annex 3 Section 2.7. Candidates with a 'low' SJE score would have to attain a higher score on the interview in order to pass the conscientiousness criterion overall.

7.8.5 Reliable and valid psychometric assessment scores

The SJE and MMI were evaluated as part of the T948 trial and both methods were revised to optimise them for the assessment of candidate train drivers. The SJE conscientiousness score was significantly correlated with relevant manager ratings of train driver

[1] Raw scores were standardised against the existing norm group so that scores have a mean of 100 and a standard deviation of 10 (producing a z100 score). This approach asissts assessors in making comparisons between the scores obtained for the various main and sub-criteria on the SJE.

conscientiousness and some safety performance measures. Many of the sub-criteria scales that make up conscientiousness also displayed significant correlations with job performance.

The MMI conscientiousness score was also significantly correlated with manager ratings of conscientiousness.

7.8.6 Each psychometric assessment score provides unique information

The SJE conscientiousness score was highly correlated with other SJE scores. This is probably because conscientiousness is a key attribute that underpins many other types of performance. All of the SJE scores would have practical value in informing the MMI and there is a strong theoretical basis for recommending all three main criteria scores despite the high correlations between them. As the SJE was only evaluated with an existing driver / trainee sample it is recommended to use all of the main criteria scores with a candidate population and to assess whether they all provide value with this population at the earliest opportunity.

The MMI conscientiousness score was highly correlated with the dealing with challenging situations score and moderately correlated with the tolerance for low stimulation score. The same rationale applies for why it is recommended to use all three MMI scores.

7.8.7 Pass marks screen out potentially unsafe candidates

The SJE is not recommended to be used with a pass mark. Low scores would be used as a trigger for a more thorough interview process.

All trial participants with job performance data were adequate performers in terms of behaviour. Therefore, the cut-off for the low SJE score bands for all three behavioural criteria was set at approximately the lowest score attained by an existing driver. The spread of scores in the candidate population is expected to be much larger.

All candidates, including those who score in the low band on the SJE, would fail the MMI if they give negative evidence of behaviour during the interview. This would provide a clear link between the interview results and safety because the appropriate behaviours that are expected of a safe train driver are clearly defined within the MMI.

7.8.8 Reasonable overall pass rate for each assessment method

The SJE and MMI were only trialled with the driver and trainee sample used during T948 and this sample was assumed to be restricted in terms of the range of scores attained. Consequently, there were no norm data for these assessment methods to indicate how performance on these assessments might be distributed in the candidate population. Therefore, it was not possible to estimate the pass rate for any of the behavioural selection criteria.

As the scoring rationale for the MMI is quite similar to the current interview process it is expected that the pass rate would be broadly in line with the current pass rate which is just below 70%.

7.9 Dealing with challenging situations

7.9.1 Definition

Dealing with challenging situations – Can exercise self-control and perform effectively when faced with difficulties, taking control of situations when necessary.

Dealing with challenging situations is broken down into the following dimensions:

Proactivity - Takes the initiative when reporting or dealing with issues. Anticipates problems and takes appropriate actions. Accepts responsibility for own actions and does not over-rely on others.

Tenacity - In the face of difficulties or pressure, has the determination and perseverance to complete a task in time and do it properly without asking for help.

Assertiveness - Confident, direct and objective in dealing with others, challenging other people when appropriate.

Calmness under pressure - In a pressured situation remains calm, shows insight into own and others' emotional reactions and takes steps to manage these.

Reactivity to stress - In a pressured situation, able to maintain effective performance (in terms of quality and rational decisions / actions).

7.9.2 Recommendation

The dealing with challenging situations selection criterion is recommended to be assessed using the SJE dealing with challenging situations score and the MMI dealing with challenging situations score as shown in Table 9.

Table 9 – Dealing with challenging situations selection criterion – Recommended assessment methods

Selection criterion	Assessment method	Score	Pass criteria
Dealing with challenging situations	SJE	Dealing with challenging situations main criterion average score	There is no minimum requirement but the score informs the MMI. A z100 score of < 77.5 = 'low' banding which requires a higher MMI mark to pass the dealing with challenging situations criterion.
	MMI	Topic area 4 score	≥ 3
		Topic area 5 score	≥ 3
		And if 'low' band on SJE:	
		Dealing with challenging situations main criterion average score	≥ 4

7.9.3 Coverage of the selection criteria

The SJE and MMI were both specifically designed to assess dealing with challenging situations according to the agreed definition and sub-criteria so the entire criterion is well covered.

7.9.4 Use of multiple independent measures

The SJE and MMI would provide two independent measures of the dealing with challenging situations criterion. The SJE would be used to inform the MMI. Candidates who score in the 'low' or 'moderate' band on the SJE would be subject to situational questions in the interview. Candidates with a 'low' score on the SJE would need to attain a higher score on the interview in order to pass the dealing with challenging situations criterion overall.

7.9.5 Reliable and valid psychometric assessment scores

The SJE and MMI were evaluated as part of the T948 trial and both methods were revised to optimise them for the assessment of candidate train drivers. The SJE dealing with challenging situations score was significantly correlated with relevant manager ratings of train driver behaviour. Some of the sub-criteria scales that make up dealing with challenging situations also displayed significant correlations with job performance.

The MMI dealing with challenging situations score did not have a significant correlation with manager ratings of relevant behaviour but

did correlate significantly with some measures of job performance. The score was judged to have excellent construct and face validity and would be used in combination with the SJE score which demonstrated strong criterion validity. On the basis of all the evidence together it is recommended for use.

7.9.6 Each psychometric assessment score provides unique information

See section 7.8.6.

7.9.7 Pass marks screen out potentially unsafe candidates

See section 7.8.7.

7.9.8 Reasonable overall pass rate for each assessment method

See section 7.8.8.

7.10 Tolerance for low stimulation

7.10.1 Definition

Tolerance for low stimulation – Capable of maintaining a good standard of performance in repetitive or monotonous work conditions.

Tolerance for low stimulation is broken down into the following dimensions:

Social need - Low need for social stimulation.

Sensation seeking - In repetitive and / or monotonous situations, able to work consistently and does not make changes on impulse.

Need for external stimulation (extraversion) - Is able to maintain performance in repetitive / monotonous situations without needing other types of stimulation to keep them alert and focused on their work.

7.10.2 Recommendation

The tolerance for low stimulation selection criterion is recommended to be assessed using the SJE tolerance for low stimulation score and the MMI tolerance for low stimulation score as shown in Table 10.

Table 10 – Tolerance for low stimulation selection criterion – Recommended assessment methods

Selection criteria	Assessment method	Score	Pass criteria
Tolerance for low stimulation	SJE	Tolerance for low stimulation main criterion average score	There is no fail but the score informs the MMI. A z100 score of < 77.5 = 'low' banding which requires a higher MMI mark to pass tolerance for low stimulation criteria.
	MMI	Topic area 6 score	≥ 3
		And if 'low' band on SJE:	
		Tolerance for low stimulation main criterion average score	≥ 4

7.10.3 Coverage of the selection criteria

The SJE and MMI were both specifically designed to assess tolerance for low stimulation according to the agreed definition and sub-criteria so the entire criterion is well covered.

7.10.4 Use of multiple independent measures

The SJE and MMI would provide two independent measures of the tolerance for low stimulation criterion. The SJE would be used to inform the MMI. Candidates who score in the 'low' or 'moderate' band on the SJE would be subject to situational questions in the interview as well as behavioural questions. Candidates with a 'low' score on the SJE would have to attain a higher score on the interview in order to pass the tolerance for low stimulation criterion overall.

7.10.5 Reliable and valid psychometric assessment scores

The SJE and MMI were evaluated as part of the T948 trial and both methods were revised to optimise them for the assessment of candidate train drivers. The SJE tolerance for low stimulation score was significantly correlated with relevant manager ratings of train driver behaviour. Some of the sub-criteria scales that make up tolerance for low stimulation had significant correlations with job performance.

The MMI tolerance for low stimulation score did not have a significant correlation with manager ratings of relevant behaviour. However, the score was judged to have excellent construct and face validity and would be used in combination with the SJE score which showed strong criterion validity. On the basis of all the evidence together it is recommended for use.

7.10.6 Each psychometric assessment score provides unique information

See section 7.8.6.

7.10.7 Pass marks screen out potentially unsafe candidates

See section 7.8.7.

7.10.8 Reasonable overall pass rate for each assessment method

See section 7.8.8.

7.11 Communication

7.11.1 Definition

The ability to read, listen, understand and respond appropriately, and effectively convey information verbally and in writing.

7.11.2 Recommendation

The MMI verbal communication score is recommended to provide a pass/fail assessment of verbal communication. The written communication test (WCT) is recommended to provide a qualitative assessment of written communication. The recommended scoring framework for the communication assessment is presented in Table 11.

Table 11 – Communication criterion – Recommended assessment method

Assessment method	Score	Raw score pass criteria
MMI	Communication	≥ 3
WCT	Legibility	No pass mark. Scores banded: = 0: Low = 1: Moderate = 2: Good Candidates who do not meet ≥ 1 (have legible handwriting) would be unable to obtain an overall moderate or good score on the WCT.
	Details section (accuracy)	No pass mark. Scores banded: ≤ 1: Low = 2 - 3: Moderate ≥ 4: Good
	Summary section (written comprehension)	No pass mark. Scores banded: ≤ 1: Low = 2-3: Moderate ≥ 4: Good
	Structure section (logical sequencing and relevance)	No pass mark. Scores banded: = 0: Low = 1: Moderate ≥ 2: Good
	Overall WCT score	No pass mark. Overall score only indicative of quality within the following bands: Good = Good on all WCT sections Moderate = Moderate score on one or more WCT section Low = Low score on one or more WCT section Illegible = Low score on legibility section regardless of banding on other sections.

7.11.3 Coverage of the selection criteria

The MMI and WCT in combination cover spoken, listening and written communication. The focus of the MMI verbal communication assessment would be on the extent to which the candidate would be understood on the assumption that they will then be capable of being trained in safety critical communications.

The WCT was specifically developed to check that candidates attain the level of written communication skills necessary for typical train drivers' written tasks.

Reading would not be specifically assessed in the new recommended process. However, several of the recommended assessment methods would require the comprehension of written content (eg TRP1) or of written instructions. So, in order to perform well in the assessment overall, a candidate would need to be able to read at a reasonable level. In addition, writing skill is closely associated with reading skill and any candidate who cannot read would be likely to perform poorly on the WCT. Therefore, a specific measure of reading ability was not considered necessary.

7.11.4 Use of multiple independent measures

Communication is recommended to be assessed by the MMI and the WCT. However, they would assess quite different aspects of communication so perhaps could not be considered to be measuring the same criteria independently.

7.11.5 Reliable and valid psychometric assessment scores

The evaluation of the MMI verbal communication score that was trialled as part of T948 highlighted several areas for improvement. The statistical evaluation of reliability and validity did not show favourable results. As a result the rating scale used for the communication measurement in the MMI was revised and is expected to provide a reliable and valid assessment if implemented. The MMI is recommended for the assessment of the other behavioural criteria and it would not be efficient to include another assessment method just for the purpose of checking verbal communication. The updated communication scale should be re-evaluated when possible to check that the improvements have had the desired effect.

The WCT was also significantly improved on the basis of the results from the T948 trials. The WCT scores provided a valid assessment of various aspects of written communication skill and was significantly correlated with manager ratings of communication skills. However, the test has several attributes which lead to it having low internal consistency, although its inter-rater reliability was good. The WCT is considered to be suitable for the purpose for which it was designed; that is to provide a quick check that candidates have a basic level of written skill. It provides a more systematic and structured assessment of written communication than the current method of using the CBI pre-interview form. However, it should not be considered to be a full psychometric assessment and it is not recommended to make pass/fail decisions.

7.11.6 Each psychometric assessment score provides unique information

The MMI verbal communication score and the WCT assess clearly different aspects of communication.

7.11.7 Pass marks screen out potentially unsafe candidates

The pass mark for verbal communication would require that the candidate can be clearly understood. Minor evidence of poor communication would be tolerated (eg the use of slang) on the assumption that communication can be improved through training. However, if too much negative evidence is displayed then the candidate would not pass. There would be a clear link to safety because candidates with very poor communication who would be hard to understand during safety critical communication would be excluded.

Written communication skill was not considered to be safety critical because most of the writing done by a train driver is done in an office context without time pressure. Therefore, the WCT is recommended for use to identify those who have a weakness in a particular area of written communication. A score banding system was proposed that would help assessors to identify candidates with poor written communication skills and to understand what aspects of written communication skill are weak. This information could be used to identify training needs. If written communication is considered by a duty holder to be critical to safe operations then individual duty holders could define their own pass mark and assess written communication as an additional pass/fail criterion.

7.11.8 Reasonable overall pass rate for each assessment method

Regarding the MMI, see section 7.8.8.

The WCT is recommended for use without a cut-off score so there is no need to estimate a pass rate. The WCT was only trialled with the driver and trainee sample used during T948 and there were no norm data for the WCT to indicate how performance on this test might be distributed in the candidate population. Therefore, it was not possible to estimate the proportion of candidates who would fall within each band.

7.12 The overall process

7.12.1 Reasonable overall pass rate for the process

It was not possible to estimate pass rates for the behavioural selection criteria because there were no data to base the estimate on. The pass rates for the cognitive and psychomotor selection criteria were estimated based on the information that was available and are presented below.

Due to the incomplete norm group data for the various assessment methods it was not possible to estimate the overall pass rate with any accuracy. Based on the information that was available estimates were made for each assessment method and each selection criterion. These estimates are conservative and probably represent an upper limit of the percentages of candidates that would be excluded.

They are summarised in Table 12 and Table 13 but are only provided for illustrative purposes and the actual pass rates will not be known until the new assessment process has been in use for some time.

Table 12 – Estimated pass rates for each cognitive and psychomotor assessment method

Assessment method	Cumulative % excluded during assessment day	Additional % of total candidate pool excluded
TEA-Occ	18	18
Paper Group Bourdon	34	16
TRP1	38	4
TRP2	41	3
WAFV	53	12
TAVTMB	56	3
2HAND	58	2

Table 13 – Estimated pass rates for each selection criterion

Selection criterion	Cumulative % excluded during assessment day	Additional % of total candidate pool excluded
Attention	34	34
Memory	38	4
Reasoning	41	3
Vigilance	50	9
Reaction time	53	3
Perception	56	3
Hand-coordination	58	2

7.12.2 Fairness

Every effort was made to collect information from females, older people and participants from minority ethnic groups. However, people from these demographic groups were so poorly represented in the train driver population that it was only possible to obtain a very small sample. In addition to trial data, test manuals were reviewed to gain further information on the fairness of the assessment methods where protected characteristics such as age, gender and ethnicity were better represented.

Assessment method scores were compared between groups. This was taken into account when proposing pass marks. Using trial data from T628 and T948, the pass rate for different groups was assessed to check whether the recommended pass marks would be likely to result in adverse impact for minorities. The benchmark for this judgement was whether the pass rate for the minority group exceeded 80% of the pass rate for the majority group (4/5^{ths} rule). The results of this are shown in Table 14. However, the small sample sizes in the minority groups meant that differences of just one person resulted in a huge increase in the percentages. The results were therefore not conclusive but are provided to show that fairness was assessed as thoroughly as possible with the available information. Appendix I contains a table with the full pass rates for each individual assessment score when the recommended scoring rules were applied to the trial data and to previous assessment data where available.

The recommended process was designed to be fair and to address the problems with adverse impact that are present in the current process. If the changes are implemented, further evaluation would be needed to confirm if this has been successful when the new process has been in use for some time and larger samples are obtained for minority groups.

Table 14 – Pass rates for T628 and T948 trial participants from different minority groups when the proposed scoring rules are applied

Predicted pass rates for trial participants from different groups							
Group		*Pass n*	*Pass %*	*Fail n*	*Fail %*	*Passes 4/5ths rule*	*Number of assessments*
Ethnicity	*White*	310	88	12	12	Yes	354
	Other ethnicity	16	70	30	30		23
Gender	*Males*	324	90	10	10	Yes	361
	Female	21	88	12	12		24
Age	*50 and under*	303	91	9	9	Yes	333
	51 and above	41	80	20	20		51

7.12.3 A note on fairness versus safety

In order to comply with the legislation on fairness and equality, one of the pass marks of the Group Bourdon and both of the pass marks for the TRP were recommended to be reduced. This might be considered at face value to be a reduction in an existing risk-control; however these adjustments can be justified from a safety perspective for the following reasons:

- All candidates would be assessed against the same pass criteria that have been set based on thorough research that had safety as the main priority.
- The current Group Bourdon pass criteria are not currently set at a 'safe' level, they are set at an unreasonable high level. They currently exclude over 40% of candidates and almost 60% of black candidates. It is highly improbable that such a high percentage of people are 'unsafe'. The validity evidence for the Group Bourdon alone was not strong enough to suggest that it distinguishes between safe and unsafe applicants with enough accuracy to justify this high exclusion rate.
- In the new recommended process the Group Bourdon would be supplemented with the TEA-Occ to provide a additional assessment of attention. The new assessment would cover selective and divided attention. The current assessment only covers selective attention.
- A vigilance assessment would also be added to further scrutinise candidates.
- The TRP tests that measure memory and reasoning were primarily related to training outcomes. The pass marks were only recommended to be relaxed by one point. This would bring it back to the level it was set at before 2007.
- The non-technical skills/behavioural assessment would be greatly enhanced and covers tolerance for low stimulation in more detail using two different assessment methods.
- The assessment process is only one small part of the overall system of risk control for train drivers. Other risk controls include training, competence management systems and engineering controls. These additional risk controls are important and are likely to make a greater contribution to safety than psychometric testing.
- Underlying cognitive ability is only one part of the reason why someone might make an error. Other individual, job and organisational factors are known to also play a role.
- No psychometric assessment process can perfectly distinguish safe and unsafe people. The process as a whole works together to provide a comprehensive assessment that increases the probability that candidates who are selected are suitable. However, in the current process, some unsuitable people slip through the net and some good people are excluded and that will continue to be the case.

The recommended changes would only be introduced after a period of consultation with industry. As a further comfort, the on-going industry strategy for train driver psychometric assessment includes

plans to evaluate the new process at the earliest opportunity and to quickly make changes if there are any concerns.

7.12.4 Assessment can be conducted in one day

The whole assessment process could be conducted in one working day as illustrated in Table 15. However, it would be longer in duration than the current assessment process because there would be an increased number of selection criteria and corresponding assessment methods. The example presented in Table 15 is just one of the options and assessment centres would have flexibility in how they administer the process.

The time estimates in Table 15 include approximate administration time but do not include time for scoring. For paper assessment methods this would be directly influenced by the number of candidates who are assessed and for computer assessments the scoring would be done automatically. A nominal amount of time for introduction/administration is included in Table 15 but this will depend on the number of candidates and the particular procedures at each assessment centre.

The number of candidates that could be assessed in one day would be dependent on the resources of the assessment centre in terms of staff and computers. For example, the number of paper tests that could be scored depends on the number of assessors and the number of interviews that could be conducted simultaneously depends on the number of qualified interviewers and available rooms.

Table 15 – Illustrative assessment day timetable

Start time	Tests	Paper/ computer	Time including approximate administration time (minutes)
09:00	Introduction and administration	n/a	20
09:20	TEA-occ	Paper	30
09:50	Paper Group Bourdon		25
10:15	TRP 1 + 2		45
11:00	30m break		
11:30	WAFV	Computer	40m
12:10	ATAVT		15m
12:25	2HAND		15m
12:40	45m lunch		
13:25	SJE	Paper and computer	45m
14:10	WCT		15m
14:25	15m break		
14:40	Interview	n/a	60m
15:40	Interview	n/a	60m

7.12.5 Ability to sift candidates based on paper assessment methods

Excluding the WCT, the recommended assessment process includes three assessment methods conducted on paper that can be used to filter large groups of candidates before the computer assessment. The time that this would take depends on the number of staff available to score the assessments. Scoring of paper based assessments would be more time consuming than computer assessments which are automatically scored.

8 Benefits of the research

Research project T948 was part of a wider programme of RSSB research work on train driver selection. Some benefits of this programme have already been realised. The formation of the Driver Selection Governance Group (described in Section 3) resulted from the review of the management arrangements that was commissioned as part of project T628 (RSSB, 2010). This provided an explicit definition of the roles and responsibilities of different organisations with respect to train driver psychometric assessment.

The group has produced a comprehensive industry strategy to support the implementation of the new recommended assessment process. The strategy covers periodic monitoring of the effectiveness of the process that will help industry achieve on-going compliance with legal requirements and continuous improvement in this area.

The main benefits of project T948 can only be realised if the recommendations are implemented via an update of RIS-3751-TOM *Train Driver Selection.* These changes would be subject to the normal standards development process which includes industry consultation and approval by the Traffic Operations and Management Standards Committee. If the recommended process is implemented, the following benefits would be expected:

- Selection criteria would be used that better cover the attributes required for modern train driving. In particular, vigilance would be introduced as a specific selection criterion. This is a benefit because it is an essential skill in light of increasing automation within the train driving task. The behavioural criteria would be enhanced and would be in line with the non-technical skills training initiatives that are currently on-going within industry (RSSB, 2010).
- The selection criteria would all be clearly described. This would assist duty holders to design their wider recruitment processes through an enhanced understanding of what the standardised psychometric assessment process does and does not include.
- The selection criteria would be compliant with the requirements of the TDLCR (2010). The language used to describe the selection criteria would be consistent with TDLCR so that compliance is obvious. The selection criteria that are currently in use do not include perception or reasoning so are not compliant.
- The new psychometric assessment process would address all of the updated selection criteria. The current assessment process does not.
- Each assessment method in the new process was thoroughly evaluated. The evaluation demonstrated the value of each assessment in measuring job relevant attributes. If implemented, duty holders can have increased confidence that the assessment will exclude candidates who might not make safe train drivers.
- The comprehensiveness of the evaluation means that the strengths and weaknesses of the new process are known. The new process would make appropriate use of each method based on this knowledge.

- A significant proportion of the recommended process would use the Vienna Test System (VTS) platform. The use of a modern computerised test platform would make the process of administration and scoring easier. The suppliers of the VTS have a strong track record of delivering psychometric assessment services to rail clients in mainland Europe. Better maintenance and support would be expected for the VTS than for the computerised assessments that are currently in use.
- The 2010/2011 RSSB review of the fairness of the current process highlighted significant concerns regarding the pass rates for ethnic minority candidates (RSSB, 2011a). The new recommended scoring rules would address these concerns and should result in a more equal pass rate for each assessment method. Furthermore, the validation evidence demonstrated the suitability of the assessment methods for the measurement of safety critical job aptitudes so if differences remain when the new process is implemented then they could be justified.
- The rail industry in GB is increasingly interested in how non-technical skills support safe performance at work, particularly in the role of train drivers. The recommended assessment process would provide an enhanced measure of the behavioural or non-technical skills required for safe train driving. It would be aligned to other work in this area and would provide a more comprehensive assessment that would be consistent with recognised good practice.

9 Considerations for duty holders

The recommendations of this research will be progressed as a proposal to update Rail Industry Standard RIS-3751-TOM Train Driver Selection Issue One. This change will be subject to the normal industry standards process. The Traffic Operations and Management (TOM) Standards Committee is responsible for making a decision about the change.

Rail Industry Standards are produced by RSSB at the request of industry where there are expected to be benefits from different companies using a common standard. Unlike Railway Group Standards, they are not mandatory. If these recommendations are implemented by inclusion in RIS-3751-TOM then each employer of train drivers would need to decide whether to use them as part of their train driver selection process.

Employers must satisfy themselves that the selection procedures they use will not illegally discriminate against any applicants and will make sure as far as reasonably practicable that only competent persons are selected. Employers of train drivers are responsible for verifying the appropriateness of their selection process and fitness

for the purpose for which it is being used. Therefore, it is important that employers understand the strengths and limitations of the research that has resulted in these recommendations. They are documented in this section.

The TDLCR (2010) imposes requirements for certain aptitudes to be assessed prior to employment as a train driver. The selection criteria recommended as a result of this research incorporate these requirements so following the recommendations of this research would provide a straightforward way for rail operators to demonstrate compliance with this aspect of legislation. The exception is the TDLCR requirement for cross-border drivers to have sufficient foreign language skills for clear safety critical communications. This is not included in the new recommended psychometric assessment process because it does not apply to the majority of operators in GB. Operators in GB who do employ cross-border drivers will need to assess foreign language skills separately.

The research that has formed the basis of the recommendations followed a robust process that took into account all available evidence. Each aspect of the recommended psychometric assessment process was subject to intense scrutiny. The evaluation criteria against which each assessment method was assessed were comprehensive. A strong rail industry steering group consisting of assessment centre staff was involved throughout all three research projects and made key decisions at each stage. In addition, RSSB's work in this area was overseen by the Driver Selection Governance Group, a sub-committee of the TOM Standards Committee.

However, any research has strengths and weaknesses. Practical and methodological limitations of this research mean that there are still some questions over how the recommended assessment process will perform if implemented. The following paragraphs describe the main limitations and the impact that they have on the conclusions of the research.

A concurrent validity research design was used because it allowed the relationship between assessment scores and on-the-job performance to be examined without having to track new drivers through their career. The existing trainees and drivers who took part in this study had already been selected on the basis of very similar selection criteria to those assessed during the research. In addition, train drivers are subject to rigorous competence management. The result is that the sample did not contain many poor performers. This restriction in the range of data affected the analysis of the correlations between assessment scores and job performance. Conclusions regarding the criterion validity of each assessment method score were based on this data and there is a chance that they would not be replicated if assessed again in future with a different sample and / or method.

During the research efforts were made to include a variety of people including failed candidates. However, job performance data were needed to assess criterion validity so the sample was predominantly made up of existing train drivers and trainees. The recommended assessment methods were not trialled on a candidate population. The estimates of what the pass rates of the assessment would be if the recommended process is implemented relied on estimating how the candidate population would perform. The VTS assessment methods all had norm data that shows how a large sample comparable to a candidate sample performed. Therefore, for VTS assessment methods the pass rate was estimated with some confidence. The TEA-Occ had limited norm data from a sample of drivers and signallers. The TEA-Occ norm data were not considered to be representative of the candidate population so their utility in the analysis was limited. The SJE and MMI did not have norm data so it was not possible to predict the pass rate. There is a chance that if the recommended process is implemented too many or too few candidates will pass and this could lead to difficulties during recruitment. For this reason, if the new process is implemented, RSSB intends to monitor the pass rate and will be prepared to adjust the cut-off scores if appropriate.

The sample used in the research was limited in terms of the numbers of females, people from ethnic minorities and older people. Every effort was made to recruit participants from these groups, including extending the duration of the trials by six months. Minimum acceptable numbers of candidates from these groups were eventually recruited so that a basic analysis of group differences could be performed. However, the low representation of these groups in the train driver population and the difficulty of releasing drivers to take part in the research meant that the sample was small and the analyses that could be performed were limited. Participants could only be grouped for analysis in a very crude way. Only two age categories and two ethnic group categories were used for most analyses. The research provided enough data to check the size of group differences and this was fed into the development of the recommended scoring rules. The recommended scoring was set carefully to provide the best balance between safety and fairness. Based on the research, it is expected that the pass rates for the recommended process would be more similar for different ethnic groups than pass rates of the current process.

The benchmark that was used to assess fairness was the 4/5ths rule which is recommended by the US Equal Opportunities Employment Commission. This rule states that differences in pass rates between groups with different protected characteristics are of concern if the pass mark of one group is less than 4/5ths or 80% of the other group. The 4/5ths rule was used in this research because it was the only objectively measurable criterion that was available to make a judgement against. However, it is important to be aware that it is not a legal definition of discrimination but rather a ‘rule of

thumb' that can be used to identify serious discrepancies. The Equal Opportunity Employment Commission is a federal agency of the United States Government. The 4/5ths rule is not a rule of UK or European law and it has not be adopted or recommended by any official body with jurisdiction in Great Britain.

On-going monitoring of the fairness of the process is planned and, if the new process is implemented, group pass-rates for the psychometric assessment process will be examined more thoroughly when enough data are available. If there are enough data then a more detailed breakdown of age and ethnic groups will be used. In addition to this, employers should continue to exercise their own legal responsibility to use fair selection processes. They should consider the fairness of their selection processes overall including how applicants are attracted to apply and what other decision making criteria are applied before and after the psychometric assessment process. The implementation of the new process in terms of how the assessment centre is run should also follow good practice to avoid disadvantaging certain groups of candidates. This includes using clear and simple instructions for all assessments and avoiding the use of time limits for assessments that do not relate to time critical attributes (Commission for Racial Equality, 1992).

Employers should note that the recommended process is intended to deselect candidates who might not have some of the aptitudes required to be safe train drivers. As such, it forms a recommended core assessment that all candidate train drivers should undertake. Employers might wish to assess candidates on additional selection criteria which are non-safety critical or adjust the pass marks to make the pass criteria more strict to select candidates who score very highly on the assessment methods. This is a matter for individual employers to decide based on their particular operations and is outside the scope of RSSB's work.

10 Recommendations for future work

If the recommended changes to RIS-3751-TOM are accepted, it will be a significant change for industry. A follow on project would be initiated and RSSB would continue to work with industry to prepare for implementation. RSSB would be supplier of three of the recommended assessment methods so processes would need to be developed for training, licencing, supply and maintenance. This work has already been planned and resources allocated in preparation. Other suppliers of psychometric assessment methods would also need to take some action.

In addition, RSSB would need to support the assessment centres in the preparations they would need to make, such as the procurement of equipment and licences, obtaining training and updating processes. Rail assessment centres are aware of these requirements via their involvement in the research steering group.

An ‘implementation working group’ consisting of RSSB and assessment centre staff has been formed ready to progress the necessary actions.

Assessment centres would also require assistance in developing detailed procedures for the assessment day. When the procedures have been developed, it is recommended that a run-through of the whole process is undertaken at one of the assessment centres. The run-through would check how well all aspects of the procedures work and provide an opportunity to make any changes before they are used with real candidates.

If the recommended assessment process is implemented, the overall pass rate and the pass rates of different groups should be checked as soon as enough data are available. The predicted pass rates presented in this report (Section 7.12.1) are only an estimate. It would be necessary to see how the assessment works in practice and to adjust the scoring rules if there are any concerns.

When enough candidate data have been collected, they should be used as the norm group for all of the assessment methods. This would allow the results of all the assessment methods to be understood in terms of how they relate to performance of the candidate population.

In addition to checking pass rates, the new process should be evaluated periodically, using similar criteria to this research. This is recommended as part of the Driver Selection Governance Group strategy (RSSB, 2011b). It is required to maintain a psychometric assessment process that effectively screens candidates from a safety perspective and is legal. This evaluation would require job performance data for candidates who have been recruited using the new assessment process so it would be possible two or three years after implementation.

In future, if any of the assessment methods need to be modified or replaced, it is recommended to trial them alongside the normal assessment centre process using real candidates. In this way, the evaluation will not rely on drivers being released from duty to take part in research and a larger sample can be collected. This method has the advantage that a norm group for the assessment can be collected at the same time. This method would require information to be collected over several years because job performance data would not be available until the recruited candidates have started work.

Regarding the assessment of attention, the T340 research (RSSB, 2005) recommended the replacement of the paper Group Bourdon. Based on the evaluation results of the alternative methods, this could not be justified so the paper Group Bourdon should be retained. However, both the paper Group Bourdon and the TEA-Occ did not demonstrate as strong criterion validity as ideal. There are also concerns around the difficulty of scoring the Group Bourdon

in its current format and over test security because there are versions of it available on the internet. In the short term, RACF and RSSB should work with the test publisher of the paper Group Bourdon to improve the test format to facilitate quicker and more accurate scoring. In the longer term, it would be beneficial to identify and trial alternative measures of attention that could be used to further improve this part of the assessment process.

11 References

Arthur D Little (2009). *Governance for train driver psychometric testing: Current arrangements and future options – Final report* (unpublished)

Bartram, D. (2010). Revision of the UK Test User standards and alignment with changes in Europe: Part 4 Some questions and answers. *Assessment & Development Matters*, 2 (2).

Bartram, D. (2008). *EFPA review model of the description and evaluation of psychological tests. Test review form and notes for reviewers.* Version 3.42 (2008). European Federation of Psychologists.

Brickenkamp, R. (1986). Handbuch apparativer Verfahren in der Psychologie. In U. Puhr. *VIGIL test manual – Vigilance*. Vienna: Schuhfried GmbH.

British Psychological Society (2006). *Design, Implementation and Evaluation of Assessment and Development Centres: Best Practice Guidelines.* Retrieved from: http://www.psychtesting.org.uk/download$.cfm?file_uuid=64962578-CF1C-D577-972D-52D28AEAD5CA&siteName=ptc.

British Psychological Society (2009). *Code of Ethics and conduct. Guidance published by the Ethics Committee of the British Psychological Society.* Retrieved from: http://www.bps.org.uk/sites/default/files/documents/code_of_ethics_and_conduct.pdf.

Campion, M., Purcell, E., and Brown, B. (1988). Structured interviewing: raising the psychometric properties of the employment interview. *Personnel Psychology*, 41, p25-42.

Commission for Racial Equality (1992). *Psychometric Tests and Racial Equality: a Guide for Employers.* Commission for Racial Equality

Commission for Racial Equality (1996). *A Fair Test? Selecting Train Drivers at British Rail.* Commission for Racial Equality

Data Protection Act 1998. (1998). London: HMSO.

Directive of the European Parliament and of the Council (EC) 2007/59/EC of 23 October 2007 on the certification of train drivers operating locomotives and trains on the railway. *Official Journal of the European Union L 315/51*

Dunlap, W. P., Burke, M. J., and Smith-Crowe, K. (2003). Accurate tests of statistical significance for rwg and average deviation interrater agreement indexes. *Journal of Applied Psychology*, 88 (2), p356-362.

Equality Act 2010. (2010). London: HMSO.

Fletcher, S. (2004). *Recruiting safe employees for safety-critical roles.* Research Report 271. Health & Safety Executive.

Flin, R., O'Connor, P., and Crichton, M. (2008). *Safety at the Sharp End: A Guide to Non-Technical Skills*. Hampshire: Ashgate Publishing Limited.

Gilbert, D.T., and Krull, D.S. (1988). Seeing less and knowing more: The benefits of perceptual ignorance. *Journal of Personality and Social Psychology*, 54, p193–202.

Gilbert, D.T., Krull, D.S., and Pelham, B.W. (1988) Of thoughts unspoken: Social inference and the self-regulation of behavior. *Journal of Personality and Social Psychology*, 55, p685–694.

Hagan, B. (2009). Reasonable adjustments. In Moody, S. (2009). *Dyslexia and employment: A guide for assessors, trainers and managers.* Chichester: Wiley-Blackwell.

International Test Commission (ITC) Guidelines on Test Use: English Version (1999). Stockholm: International Test Commission.

RSSB. (2005). *T340: Psychometric Testing – A Review of the Train Driver Selection Process – Deliverable 7: Validation study of the current recruitment process and review of the future train driving role.* London: RSSB.

Karner, T., and Neuwirth, W. (2000). Validation of traffic psychology tests by comparing with actual driving. In U. Puhr. *2HAND test manual.* Vienna: Schuhfried GmbH.

Latham, G.P., and Sue-Chan, C. (1999). A meta-analysis of the situational interview: an enumerative review of reasons for its validity. *Canadian Psychology* (40), p56-67.

Lievens, F., Peeters, H., and Schollaert, E. (2006). Situational judgment tests: a review of recent research. *Personnel Review*, 37(4), p426-441.

Levashina, J., and Campion, M. A. (2006). A Model of Faking Likelihood in the Employment Interview. *International journal of selection and assessment* 14 (4), p299-316.

London Fire Brigade (2008). *Policy on dyslexia and other specific learning disabilities.*

Mackworth, N.H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, 1, p6-21.

Mackworth, N.H. (1970). Vigilance and attention. In U. Puhr. *VIGIL test manual – Vigilance*. Vienna: Schuhfried GmbH.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., and Grubb, W. L. (2007). Situational judgement tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, 60, p63-91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., and Braverman, E. P. (2001). Use of situational judgement tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, 80, p730-740.

McDaniel, M. A., and Nguyen, N. T. (2001). Situational Judgement Tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, 9, p103-113.

McDaniel, M.A., Whetzel, D. L., Hartman, N. S., Nguyen, N., and Grubb, W. L. (2006). Situational judgement tests: validity and an integrative model. In R Polyhart and J Weekley (eds). *Situational judgement tests: theory, measurement and application.* Jossey Bass. p183–204.

Moody, S. (2009). *Dyslexia and employment: A guide for assessors, trainers and managers*. Chichester: Wiley-Blackwell.

Motowidlo, S. J., Hooper, A. C., and Jackson, H. L. (2006). A theoretical basis for situational judgement tests. In R Polyhart and J Weekley (eds). *Situational judgement tests: theory, measurement and application.* Jossey Bass. p57–81.

Puhr, U. (2004). *VIGIL test manual – Vigilance*. Vienna: Schuhfried GmbH.

Puhr, U. (2003). *Two-hand coordination. 2HAND test manual*. Vienna: Schuhfried GmbH.

Robertson, I.H., and Garavan, H. (2004). *Vigilant Attention. The Cognitive Neurosciences,* Ed. Gazzaniga, M. 3rd Edition.

Roth, P.L., Bobko, P., and Switzer, F.S. (2006). Modelling the Behavior of the 4/5ths Rule for Determining Adverse Impact: Reasons for Caution. *Journal of Applied Psychology* Copyright 2006 by the American Psychological Association 2006, Vol. 91, No. 3, p507–522.

RSSB (2005). *T148 Human factors associatd with driver error and violation.* London: RSSB.

RSSB (2010). *T628 Driver Selection: Development Phase - Updated selection criteria and validation study.* London: RSSB.

RSSB (2011a). *Health check review of the train driver assessment process.* Commissioned by the Driver Selection Governance Group. London: RSSB.

RSSB (2011b). *Driver Selection Governance Group Strategic Plan 2011 – 2015 Issue One.* London: RSSB.

RSSB (2012). *Non-technical skills for rail: Developing an integrated approach to NTS training and development.* London: RSSB.

RSSB (in press). *SJE and MMI manual.* London: RSSB.

Sturm, W. (2006). *WAFV test manual - Perception and attention functions: vigilance.* Vienna: Schuhfried GmbH.

Sturm & Büssing (1990). In Puhr, U. (2004). *VIGIL test manual – Vigilance.* Vienna: Schuhfried GmbH.

Taylor, P., and Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology*, 75, p277–294.

Train Driver Licences and Certificates Regulations (2010). London: HMSO

Weekley, J. A., and Polyhart, R. E. (Eds.) (2006). *Situational Judgement Tests*. London: Lawrence Erlbaum.

12 Glossary

Abbreviation/ term	**Explanation**
2HAND	The Two-Hand Coordination is a computerised psychological assessment tool developed by Schuhfried GmbH. It is designed to measure hand coordination.
ATAVT	The Adaptive Tachistoscopic Traffic Perception Test is an adaptive computerised psychological assessment tool developed by Schuhfried GmbH. It is designed to measure perception.
Behavioural indicators	Examples of behaviour that indicate performance
Group Bourdon	The Group Bourdon is a paper-based psychometric test distributed by Southeastern. It is designed to measure attention.
CBI	The Criterion Based Interview is an interview process developed by OPC. It is designed to measure the ability to communicate verbally and in writing and the following behavioural criteria; follows set rules and procedures, conscientiously works to meet training and job demands, remain calm in emergency and stressful situations, proactive and tenacious,

	and can spend time alone and does so effectively.
CIF	Candidate interview form (candidate completes the CIF at the assessment centre in advance of the CBI, and answers are used in the CBI).
ComPACT	ComPACTis a test battery developed by Hogrefe Ltd which comprises a test of vigilance.
Cognitrone	Cognitrone is a computerised assessment tool developed by Schuhfried GmbH. It is designed to measure attention and concentration.
DAUF	The Continuous Attention Test (DAUF) is a computerised assessment tool developed by Schuhfried GmbH. It is designed to measure attention.
DAKT	The Differential Attention Test is a computerised assessment tool developed by Schuhfried GmbH. It is designed to measure attention.
DCS	Dealing with challenging situations (one of the three main behavioural selection criteria)
Driver Selection Governance Group (DSGG)	An industry sub-committee of the Traffic Operations and Management Sub-Committee who have a remit to recommend industry strategy for psychometric assessment of train drivers for selection.
DTG	The Determinations Gerat test is designed to assess the operation of hand and foot controls for train driver selection.
Fairness	A test is considered to be fair if it does not unjustifiably discriminate against any respondent groups based on gender, age, ethnic origin or disability.
ITC	International Test Commission. The ITC produced guidelines for good test use and encourages best practice in assessment.
MICROPAT	MICROPAT is a simulation based assessment method developed by the UK Armed Forces to assess attention in pilots.

MMI	The Multi-Modal Interview is an interview process developed by RSSB. It is designed to be used with the Situational Judgement Exercise. It is designed to measure verbal communication and the following behavioural criteria conscientiousness, dealing with challenging situations, and tolerance for low stimulation (including the underpinning sub-criteria).
NEO-PI-R	An internationally-recognised and highly regarded questionnaire measure of the five major domains of personality published by Hogrefe
OPC	OPC Assessment Ltd is a leading provider of psychometric tests based in the UK.
OTMR/OTDR	On Train Monitoring Recorder/On Train Data Recorder. A device that records data about the operation of train controls and performance in response to those controls and other train control systems.
PIF	Pre-interview form (candidate completes the PIF at home in preparation for the MMI, and answers are used in the MMI)
PR	Percentile ranks are commonly used to clarify the interpretation of scores on standardised tests. The percentile rank of a score is the percentage of scores in its frequency distribution that are the same or lower than it. For example, a test score that is greater than 75% of the scores of people taking the test is said to be at the 75th percentile.
RACF	Rail Assessment Centre Forum – A group of representatives from Railway Assessment Centres who work together to maintain consistent and high quality administration of the psychometric assessment process defined in RIS-3751-TOM.
Rail Industry Standards (RIS)	Standards produced by RSSB at the request of industry where there are expected to be benefits from different companies using a common standard. They can be adopted as company standards.

Railway Group Standards	Documents produced pursuant to the Railway Group Standards Code (or equivalent predecessor documents) defining mandatory requirements in respect of the mainline railway.
Raw score	A test score that has not been standardised against any samples of people who have taken the test.
SCAAT	The Safe Concentration and Attention Test (SCAAT) is a test of attention developed by Assessment Ltd.
Schuhfried GmbH	The Schuhfried company is an internationally recognised leader in computer-based psychological assessment based in Vienna, Austria.
SIMKAP	The SIMKAP Simultaneous Capacity/Multi-Tasking is a computerised psychological assessment tool developed by Schuhfried GmbH. It is designed to assess simultaneous capacity and stress tolerance.
SJE	The Situational Judgement Exercise is a computerised measure of behavioural preference developed by RSSB. It is designed to be used with the Multi-Modal Interview. It is designed to measure the following behavioural criteria conscientiousness, dealing with challenging situations, tolerance for low stimulation and the underpinning sub-criteria.
SME	Subject matter expert
TAVTMB	The Tachistoscopic Traffic test (TAVTMB) is part of the computerised Vienna Test System battery. It assesses visual perception and perceptive speed in traffic situations.
TDLCR	Train Driver Licences and Certificates Regulations (2010). A piece of legislation that brings into force the requirements of European Commission Directive 2007/59/EC (on the certification of train drivers) in Great Britain. It introduces a licensing and certification system for some train drivers in Great Britain. http://www.rail-reg.gov.uk/server/show/nav.2447

TEA	The Test of Everyday Attention is a paper-based psychometric test developed by Pearson Assessments. It is designed to measure attention and concentration.
TEA-Occ	The Test of Everyday Attention for Occupational Assessment is an adapted version of the original TEA developed by Pearson Assessments for use in the rail industry.
TLS	Tolerance for low stimulation (one of the three main behavioural selection criteria)
TOAV	The Test of attentional vigilance (TOAV) is a test of attention developed by PEBL.
TOVA	The Test of variables of attention (TOVA) is a test of attention developed by the TOVA Company.
TRP	The Trainability for Rules & Procedures Test is a paper-based psychometric test developed by OPC Assessment Ltd. It is designed to measure trainability, memory and reasoning.
VTS	The Vienna Test System is a leading computerised psychological assessment tool developed by Schuhfried GmbH. The tests comply with the International Test Commission guidelines and the Standards for Educational and Psychological Testing of the American Educational Research Association.
VIGIL	The VIGIL Vigilance Test is a computerised psychological assessment tool developed by Schuhfried GmbH. It is designed to measure vigilance.
WAFV	The WAFV Vigilance Test is a computerised psychological assessment tool developed by Schuhfried GmbH. It is designed to measure vigilance.
WCT	The Written Communications Test is a paper-and-pencil test developed by RSSB to measure the ability to communicate effectively in writing.
ZBA	The Time Movement Anticipation test (ZBA) is

	part of the computerised Vienna Test System battery. It is designed to assess the anticipation of events based on speed, distances and directions.
Z100 score	A score that has been standardised based on the trial sample norm group so that the mean equals 100 and the standard deviation equals 10. A standardised score puts a test score into the context of a group of people who have already taken the test.

Annex 1 – The recommended selection criteria and definitions

Selection criteria	Definitions and sub-criteria
Attention	**Selective attention** - The ability to differentiate between different sources of information and attend selectively to them, eg distinguishing and attending to alarms (selective attention). **Divided attention** - The ability to switch attention between sources of information, eg lineside information and in-cab information and perform different tasks in parallel; making train announcements while on the move.
Vigilance	The ability to attend and respond to stimuli which occur relatively infrequently and over extended periods of time.
Memory	The ability to learn, recall and apply job related information in appropriate time limits, eg learn new information in training; remembering instructions from signallers; applying specific rules and procedures.
Reasoning	The ability to solve problems and make decisions, eg fault diagnosis; understanding and interpreting information from instrumentation.
Perception	The ability to anticipate elements in a traffic environment and make a correct decision about how to respond appropriately, eg identifying a landmark cue before a station and starting to decelerate.
Reaction time	A quick and adequate response to simple and complex visual and acoustic stimuli and the associated quality of performance.
Hand coordination	The ability to make appropriate and controlled movements in response to decisions about complex stimuli.
Communication	The ability to read, listen, understand and respond appropriately, and effectively convey information verbally and in writing.

Selection criteria	Definitions and sub-criteria
Conscientiousness	Has the drive and willingness to achieve the goals they are set and to complete work to the highest possible standard, working effectively with others as required: **Dependability** Can be relied upon to carry out the tasks and fulfil the responsibilities expected of them. **Attitude to work and people** Considerate, supportive and co-operative towards others. **Commitment to work** Does a task well and to best of ability, prioritising work over other commitments as appropriate. **Attention to detail** Is thorough in accomplishing a task and pays close attention to detail. Takes a systematic and unhurried approach. **Ability to check and not make assumptions** Checks own understanding of all relevant information, and others' understanding as appropriate. **Compliance with rules and procedures** Follows rules and procedures, understands their relevance and takes action if others do not follow rules or if rules are inappropriate.
Dealing with challenging situations (DCS)	Can exercise self-control and perform effectively when faced with difficulties, taking control of situations when necessary: **Proactivity** Takes the initiative when reporting or dealing with issues. Anticipates problems and takes appropriate actions. Accepts responsibility for own actions and does not over-rely on others. **Tenacity** In the face of difficulties or pressure, has the determination and perseverance to complete a task in time and do it properly without asking for help. **Assertiveness** Confident, direct and objective in dealing with others, challenging other people when appropriate. **Calmness under pressure** - In a pressured situation remains calm, shows insight into own and others' emotional reactions and takes steps to manage these. **Reactivity to stress** - In a pressured situation, able to maintain effective performance (in terms of quality and rational decisions/ actions).

Selection criteria	Definitions and sub-criteria
Tolerance for low stimulation (TLS)	Capable of maintaining a good standard of performance in repetitive or monotonous work conditions: **Social need** Low need for social stimulation **Sensation seeking** In repetitive and / or monotonous situations, able to work consistently and does not make changes on impulse. **Need for external stimulation (extraversion)** Is able to maintain performance in repetitive / monotonous situations without seeking to add further stimulation.

Table 16 - The recommended selection criteria and definitions

Annex 2 – The recommended psychometric assessment process and scoring rules

Important note: Some selection criteria have compensatory scoring rules whereby certain test scores are used together in a specific way to reach a pass/fail decision. Compensatory scoring rules are highlighted in grey.

Table 17 – Cognitive and psychomotor assessment process and scoring rules

Selection criterion	Assessment method	Test form/ subtests	Scoring variable(s)	Pass criteria	Scoring rule(s)
Attention	TEA-Occ	Lift counting with distraction Telephone search Telephone search while counting	Lift counting with distraction	≥ 6	Candidate must meet minimum standard on **each test score** in order to pass 'attention'
			Dual task decrement	≤ 4.44	
	Paper Group Bourdon	n/a	Total production	≥ 938	
			Total omissions	≤ 47	
Memory	TRP1	n/a	TRP1	≥ 9	Candidate must meet minimum standard in order to pass 'memory'
Reasoning	TRP2	n/a	TRP2	≥ 13	Candidate must meet minimum standard in order to pass 'reasoning'
Vigilance	WAFV	S2	Missed reactions	≤ 5	Candidate must meet minimum standard on **each test score** in order to pass 'vigilance'
			False alarms	≤ 8	

Selection criterion	Assessment method	Test form/ subtests	Scoring variable(s)	Pass criteria	Scoring rule(s)
Reaction time	WAFV	S2	Reaction time	< 656	Candidate must meet minimum standard on **each test score** in order to pass 'reaction time'
Perception	ATAVT	S2 (for countries that drive on the left)	Overview	> -1.5066	Candidate must meet minimum standard in order to pass 'perception'
Hand coordination	2HAND	S1	Overall mean duration	< 52.8	Candidate must meet minimum standard on **each test score** in order to pass 'hand coordination'
			Percent error duration	< 16.7	

Table 18 – Behavioural assessment process and scoring rules

Selection criterion	Assessment method	Test form/ subtests	Scoring variable(s)	Pass criteria	Scoring rule(s)
Conscientiousness	SJE	Version C	SJE Conscientiousness	No fail	Two bands: 'Low' (z100 score ≤ 77.49) 'Moderate/Good' (z100 score ≥ 77.5)

	MMI	n/a	Topic area 1 Topic area 2 Topic area 3	≥ 3	Candidate must meet minimum standard on **each topic area** in order to pass this part of the MMI
	MMI and SJE used to produce overall 'Conscientiousness' score		SJE Conscientiousness MMI Conscientiousness (average of topic areas 1, 2 and 3)	MMI Pass + if 'low' SJE, MMI Conscientiousness ≥ 4	Candidate must meet minimum standard on **MMI** in order to pass 'Conscientiousness' and, **if candidate scores 'low' SJE Conscientiousness**, must score ≥ 4 on MMI Conscientiousness
Dealing with challenging situations (DCS)	SJE	Version C	SJE DCS	No fail	Two bands: 'Low' (z100 score ≤ 77.49) 'Moderate/Good' (z100 score ≥ 77.5)
	MMI	n/a	Topic area 4 Topic area 5	≥ 3	Candidate must meet minimum standard on **each topic area** in order to pass this part of the MMI

	MMI and SJE used to produce overall 'DCS' score		SJE DCS MMI DCS (average of topic areas 4 and 5)	MMI Pass + if 'low' SJE, MMI DCS ≥ 4	Candidate must meet minimum standard on **MMI** in order to pass 'DCS' and, **if candidate scores 'low' SJE DCS**, must score ≥ 4 on MMI DCS
Tolerance for low stimulation (TLS)	SJE	Version C	SJE TLS	No fail	Two bands: 'Low' (z100 score ≤ 77.49) 'Moderate/Good' (z100 score ≥ 77.5)
	MMI	n/a	Topic area 6	≥ 3	Candidate must meet minimum standard on **topic area 6** in order to pass this part of the MMI
	MMI and SJE used to produce overall 'TLS' score		SJE TLS MMI TLS (topic area 6)	MMI Pass + if 'low' SJE, MMI TLS ≥ 4	Candidate must meet minimum standard on **MMI** in order to pass 'TLS' and, **if candidate scores 'low' SJE TLS**, must score ≥ 4 on MMI TLS

Table 19 – Communication assessment process and scoring rules

Selection criterion	Assessment method	Test form/ subtests	Scoring variable(s)	Pass criteria	Scoring rule(s)	
Communication	MMI	n/a	MMI Verbal Communication	≥ 3	Candidate must meet minimum standard in order to pass MMI communication	
	WCT	Version 1 and 2	Legibility	No fail	3 bands: = 0: Low = 1: Moderate = 2: Good Candidates who do not meet ≥ 1 (have legible handwriting) would be unable to obtain an overall moderate or good score on the WCT.	
			Accuracy	No fail.	3 bands: ≤ 1: Low = 2-3: Moderate ≥ 4: Good	If a candidate does not reach the 'good' level on these sections then a qualitative comment is added to the scoring sheet highlighting the weak area(s).
			Written comprehension	No fail.	3 bands: ≤ 1: Low = 2-3: Moderate ≥ 4: Good	
			Structure (consisting of logical sequencing and relevance)	No fail.	3 bands: = 0: Low = 1: Moderate ≥ 2: Good	

			Overall written communication score	No fail.	4 bands: Good = Candidate achieved good bands on all WCT sections. Moderate = Candidate achieved moderate band scores on one or more of the WCT sections. Low = Candidate achieved a low score on one or more of the WCT sections. Illegible = Candidate scored low on legibility, regardless of the score on other sections.

Annex 3 – The T948 validation study

1 Introduction

This annex report presents in detail the research conducted during project T948 *Train Driver Selection Implementation Phase.* The objectives of T948 were to:

- Evaluate the suitability of the proposed methods for measurement of behavioural criteria, communication, hand coordination and vigilance, in terms of validity, reliability, fairness and practicality.
- Propose a revised assessment centre process which will effectively measure the updated selection criteria, conform to good practice in selection, exclude unsuitable drivers and where possible enable potentially exceptional drivers to be identified.

The main activity in project T948 was an evaluation trial of the proposed assessment methods for hand coordination, vigilance, written communication and behavioural criteria. Further evaluation evidence regarding the assessment methods currently in use was also gathered to update the conclusions from T340 (RSSB, 2005). Evidence from project T340, T628 and T948 was collated in order to design a new recommended train driver assessment process.

Section 2 describes the assessment methods that were included in the T948 trial.

Section 3 outlines the criteria that were used to evaluate all assessment methods to judge whether and how they should be included in the recommended assessment process.

Section 4 sets out the method for the evaluation trial.

Section 5 presents the results of the evaluation trial for each assessment method against each of the evaluation criteria.

Section 6 considers the merits of the current assessment methods in light of the evidence from T340 and other evidence gathered more recently.

Section 7 covers the process of building a recommended assessment process and scoring rules from all the available options.

The appendices to this annex report provide further detailed information regarding the development process of some of the assessment methods and the full statistics from the evaluation trial.

2 Description of the assessment methods included in the trial

2.1 Introduction

The following methods were included in the T948 trial in order to assess their effectiveness as measures of the following criteria:

Table 20 - Selection criteria with corresponding assessment methods trialled

Selection criterion	Assessment methods trialled
Vigilance	WAFV Vigilance Test VIGIL Vigilance Test
Hand coordination	Two-Hand Coordination
Conscientiousness	Situational Judgement Exercise (SJE) Multi-Modal Interview (MMI)
Dealing with challenging situations	
Tolerance for low stimulation	
Communication	MMI Written Communication Test (WCT)

This section provides a description of each of the methods that were trialled during T948. Full details about the main scores produced by each assessment method are provided in Appendix E.

New assessment methods to cover the other selection criteria were previously trialled as part of project T628 *Train Driver Selection Development Phase* (RSSB, 2010). The assessment methods trialled were: Test of Everyday Attention for Occupational Assessment (TEA-Occ); Tachistoscopic Traffic Test (TAVTMB); Simultaneous Capacity and Stress Tolerance Test (SIMKAP) and the Time Movement Anticipation test (ZBA). These methods are described in the report for T628.

Current assessment methods were evaluated during project T340 and more recently during project T948 based on further evidence gathered from a variety of sources (see Section 6). A description of the current methods is available in the report for T340 (RSSB, 2005).

2.2 WAFV

2.2.1 What is it designed to measure?

The WAF test battery consists of six tests which assess various sub-functions of attention (Sturm, 2006). The WAFV is one of these tests and can be administered independently to assess vigilance or sustained attention. When considering the longer term application of attention, a distinction should be made between sustained attention tasks and vigilance tasks. The definition of vigilance which has been agreed with industry and will be proposed to be included

in RIS-3751-TOM is 'the ability to attend and respond to stimuli which occur relatively infrequently and over extended periods of time'. Whereas both sustained attention and vigilance make demands on attentional resources over a long period of time, the frequency and intervals of stimuli differ. Vigilance is a special type of long term attention where the stimuli typically occur at very irregular intervals and at a very low frequency compared to the number of irrelevant stimuli (Sturm, 2006). A typical vigilance task would be where an air traffic controller has to be attentive over a long period of time in order to detect a signal on the screen that stands out against irrelevant background stimuli (Mackworth, 1948). This can also be compared to the train driving task, where drivers must respond to cautionary signals or other hazards which occur infrequently. This is different to sustained attention tasks where the stimuli tend to occur in relatively higher frequencies. The WAFV was therefore chosen as a possible measure of the visual aspects of vigilant attention.

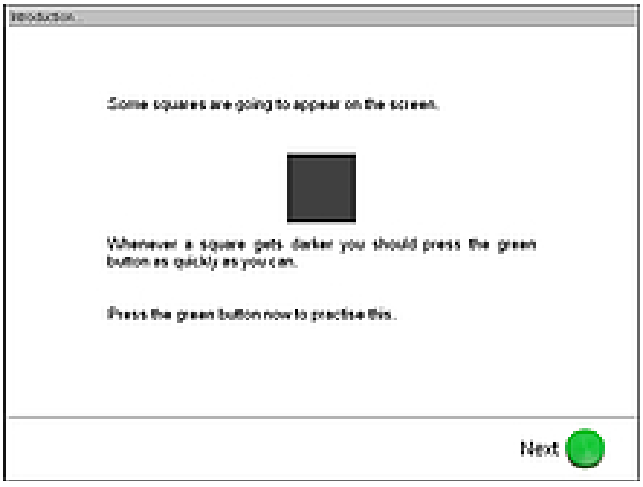
2.2.2 What does it look like?

The WAFV is a computer-administered test. All instructions are presented on the screen with an opportunity for the candidate to undertake practice tests.

Test form S2 was selected for trial as the presentation of visual stimuli (as opposed to auditory stimuli) was considered by the stakeholder group to be more relevant to the train driving role, particularly responding to visual warning signals. Further information about WAFV and the test forms is described in Appendix A.

In test form S2, the candidate is presented with visual stimuli in the form of a grey square in the centre of the screen. Occasionally, the square becomes darker in colour. The candidate's task is to respond to these occasional cases; in the case of vigilance these cases constitute 5% of the stimuli.

Figure 1 - Screen shot of the WAFV test instructions



2.2.3 Description of main variables

In WAFV, the reaction times and the various error types are scored.

- Number of missed reactions - The number of stimuli to which no response was made within 1500 ms.
- Mean reaction time - Mean of the reaction times to each target stimulus.
- Number of false alarms - Number of times a reaction key was pressed in response to irrelevant stimuli or when no stimulus had been presented.

2.3 VIGIL

2.3.1 What is it designed to measure?

Like the WAFV vigilance test, the VIGIL test is also designed to assess vigilance over extended periods of time with relatively infrequent stimuli. The test is based on the original vigilance studies by Mackworth (1970) where respondents had to watch a clock dial with a hand moving in regular jumps. Occasionally, and at irregular intervals, the hand moved twice the usual distance, which was the signal for respondents to press a button. In this study, it was found that as the observation time reached 30 minutes, the number of double jumps which the respondents failed to react to grew substantially. The research literature considers such tests as realistic measures of vigilance where the person must undertake 'monotonous monitoring tasks' (Brickenkamp, 1986). The reason these tests are useful for assessing vigilance is that performance tends to deteriorate due to a decreasing activation level of the respondent and the corresponding increase in reaction latency (Puhr, 2004). These conditions are similar to the role of driving where drivers must respond to cautionary signals which occur infrequently.

2.3.2 What does it look like?

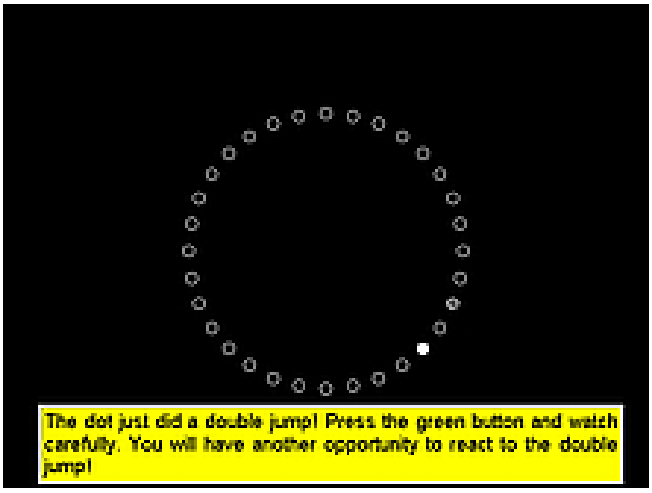
VIGIL is a computerised-administered test. All instructions are presented on the screen with an opportunity for the candidate to undertake practice tests.

Test form S2 Müggenburg 33 was selected for trial as this test form is suitable for people with normal attention al capacity. Test form S1 was not chosen as it is more appropriate for people with known attentional deficiencies. Test form S4 was not chosen as it is 70 minutes long which was considered too lengthy to fit reasonably into an assessment centre day. Further information about VIGIL and the test forms is given in Appendix A.

In test form S2, a brightly flashing dot travels along a circular path in small jumps. Occasionally, the dot takes a double jump which the candidate must respond to by pressing a button on the control panel. In the chosen test form, the outline of the circular path is not

shown; instead, the candidate must monitor the movement of the dots only.

Figure 2 - VIGIL screen shot



2.3.3 Description of main variables

- Number correct – Number of correct reactions to a double jump where the response button is pressed before the next jump takes place.
- Number incorrect – Number of responses made when no critical stimulus has occurred.
- Mean value of reaction time correct (sec) - Average time that elapses between presentation of a critical stimulus and correct response.

2.4 2HAND

2.4.1 What is it designed to measure?

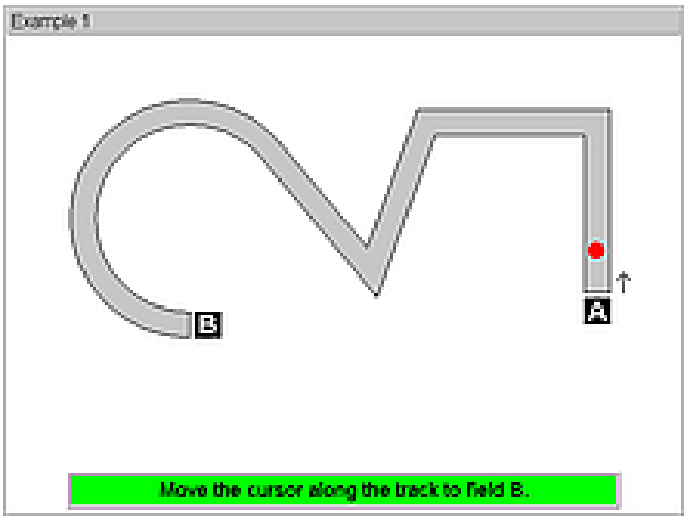
The Two-Hand Coordination test (2HAND) is designed to assess the speed and accuracy of sensory-motor coordination between the eye and hand as well as coordination between both hands.

2.4.2 What does it look like?

The 2HAND is a computer-administered test. The candidate is required to make a red dot move along a given track using two joysticks. The track consists of three sections that make increasingly difficult demands on the coordination of the left and right hand. The point is moved from right to left.

Test form S1 was selected for trial as joysticks were considered more suitable than control knobs for assessing hand-hand coordination. Further information about 2HAND and the test forms is described in Appendix A.

Figure 3 - 2HAND screen shot



2.4.3 Description of main variables

- Total mean duration - The average time needed to run through the track.
- Total mean error duration - The average time during which the red dot is outside the parameters of the track.
- Total percent error duration – The average percentage of the total duration that the red dot is outside the parameters of the track.
- Coordination difficulty – A measure of the coordination performance of the respondent calculated by a comparison between performance when two hand coordination is required and when it is not.

2.5 WCT

2.5.1 What is it designed to measure?

The WCT was developed to assess whether candidates have the level of written communication skills necessary for typical train driver written tasks. The WCT was developed by RSSB, in consultation with a psychometric testing expert, an industry steering group and a designer. The test gathers evidence on four areas of written communication: accuracy, comprehension, legibility and structure. For more information on the approach taken to develop the WCT and the associated rationale, please refer to refer to Appendix B.

The WCT is intended to be used in conjunction with the verbal and listening skills assessed in the MMI to contribute to the overall assessment of communication.

2.5.2 What does it look like?

The WCT is a short ‘paper and pencil’ test that consists of a six-picture storyboard and other details based on a transport theme that candidates must write about on a report form.

There are two versions of the WCT with different storyboards on each. Candidates are randomly assigned either one of the versions during the assessment.

After the candidates have completed the test, assessors mark the report form by hand. Points are awarded using a scoring framework. The report form provides a score for each sub category and a total score of written communication which is banded into three bands: 'low', 'needs development' and 'acceptable' levels of written communication.

2.5.3 Description of variables

The WCT has been designed to provide scores on four areas of written communication:

- Accuracy - Key details are reproduced accurately and without error.
- Written comprehension - Key concepts within the storyboard are described in a meaningful way.
- Legibility - The writing is clear to read.
- Structure - The information is presented in a logical sequence and is relevant to the story being described.

The final versions of the WCT differ from the trialled versions because some changes were made to the scoring framework to increase score variance. There were also a few changes made to the storyboards, instructions pages and report forms. Details about the rationale and the developments that led to these decisions can be found in Appendix B.

2.6 SJE

2.6.1 What it is designed to measure?

The Situational Judgement Exercise (SJE) was designed to gather evidence on behavioural preferences in relation to each of the behavioural main and sub- criteria (see Annex 1).

The SJE has been developed for use before the Multi-Modal Interview in order to target and focus the interview. The SJE is not intended as a pass/ fail measure; the results are used to highlight any potential weaker areas which are then explored in greater depth in the interview.

The SJE was developed in-house by RSSB, in consultation with a psychometric testing expert and an industry steering group. For more information on the approach taken to development, and the rationale behind development decisions, please refer to Appendix C.

2.6.2 What does it look like?

The SJE is a computer-administered exercise that consists of hypothetical scenarios, each with up to six response options. The

response options are actions that could be taken in response to the scenario, and the candidate is required to rate each action.

The scenarios are written so as to present a dilemma, so there is not always a clear correct response. Sometimes, in selecting a rating, a candidate may score highly on one sub-criterion to the detriment of another.

Once the candidate has completed the exercise, a report is made available for the administrator to download. This report charts the candidate's performance for each of the main criteria, and provides ratings for each sub-criterion, as well as information on the consistency of responses. This report doubles as an interview schedule for the interviewer conducting the Multi-Modal Interview (MMI), highlighting how scores attained on the SJE should determine aspects of the interview.

Note that both the trialled versions of the SJE (version A and version B) consisted of two different types of rating scale – a 1-5 scale of how helpful or unhelpful the action is, and a 1-4 scale of how likely or unlikely the candidate would be to take that action. Version B was found to be less effective than version A and so the trial results in Section 5 of this Annex report relate to SJE version A.

Following the trials, it was decided that only the helpful / unhelpful rating scale should be used. There was also a change in the number of scenarios so that the final version of the SJE has 21 scenarios. Details of these development decisions are provided in Appendix C.

2.6.3 Description of variables

2.6.3.1 Main criteria scores

The SJE has been designed to provide a score on each of the behavioural sub-criteria, which are then averaged to produce a score for each of the three main selection criteria (Conscientiousness, DCS, and TLS). The score indicates how similar the candidate's responses were to the ideal train driver response.

As well as providing raw scores for the three main criteria, all scores (main and sub) are then standardised against the existing norm group so that scores have a mean of 100 and a standard deviation of 10[2]. Main criteria scores are banded into one of two bands; 'low' or 'moderate/good'.

2.6.3.2 Sub-criteria scores

Information is also collected on the sub-criteria scores and consistency in order to guide the MMI process.

[2] Note that because the SJE has not yet been implemented outside the trials the existing norm group is restricted to the trial sample. Extensive norm group data will be collected when the SJE has been in use with candidate populations.

Sub-criteria scores are standardised against the existing norm group[1] so that scores have a mean of 100 and a standard deviation of 10. Sub-criteria scores are banded into one of three bands; 'low', 'moderate' or 'good'.

Information is also collected on the consistency of responses across all items relating to each sub-criterion ie did the candidate provide responses that were consistently similar / dissimilar to the ideal train driver response or was there a range in the responses given. For each sub-criterion consistency information is banded into 'weak', 'medium' or 'strong'.

Further information on how these scores are used to guide the MMI is provided in Annex 3 Section 7.4.

For further information on the SJE please refer to the SJE and MMI manual.

2.7 MMI

2.7.1 What it is designed to measure?

The MMI was specifically developed to be used with the results of the SJE to make a judgement about the extent to which the candidate's behavioural preferences satisfy each of the behavioural selection criteria.

The MMI was designed to take into account both what the candidate has done in the past, and how they might behave in a future situation. It is also designed to be a simple measure of verbal communication.

2.7.2 What does it look like?

The final version of the MMI is divided into six topic areas, each of which are designed to cover one or more of the selection sub-criteria:

- Three topic areas cover the six sub-criteria related to 'Conscientiousness'.
- Two topic areas cover the five sub-criteria related to 'Dealing with challenging situations (DCS)'.
- A final topic area covers the 'Tolerance for low stimulation (TLS)' sub-criteria.

Before the interview, candidates complete a pre-interview form (PIF). This is structured around the same topic areas, and instructs candidates to provide a brief answer to each question about their experience in this area. The candidate can provide examples from their home, social, educational or working life. They are instructed that if they cannot think of an example they should leave the box blank. The candidate completes the PIF before attending the assessment centre.

In the interview, a behavioural question (ie a question about how the candidate has behaved in the past) is used for each of the six topic

areas. This question is formulated from the information provided on the pre-interview form.

For each topic area, a situational question (ie a question about how the candidate would respond in a hypothetical scenario) is used in addition to the default behavioural question when:

- The candidate receives a 'low' or 'moderate' score on the SJE for any of the sub-criteria in that topic area.
- The consistency of the candidate's SJE scores for any of the sub-criteria in that topic area is 'weak'.
- The candidate is unable to provide a behavioural example for that topic area.

2.7.3 Scoring

Behavioural indicators have been developed for each sub-criterion, and in scoring the candidate's responses to the topic area the interviewer takes the coverage of these indicators into account.

The interviewer also rates the candidates' verbal communication skills at the end of the interview using a scale reflecting the proportion and quality of behavioural indicators.

2.7.4 Description of main variables

- Main selection criteria scores - The MMI has been designed to provide a 1-5 score for each of three main selection criteria. This is calculated by averaging scores attained on the related topic areas (topics areas 1, 2, and 3 relate to Conscientiousness, 4 and 5 to DCS and 6 to TLS). Topic area scores are based on the proportion of positive and negative indicators evidenced for the relevant sub-criteria.
- Verbal communication score - The MMI provides a verbal communication score which is on a scale of 1-5.

For further information on the MMI, including the topic areas and behavioural indicators please refer to the SJE and MMI manual.

It should be noted that the final version of the MMI differs from the trialled version and that the trial results in Annex 3 Section 5 relate to the trialled version. The trialled version:

- Consisted of eight rather than six topic areas. As with the final version, each topic area related to one or more sub-criteria.
- Collected individual sub-criterion scores, rather than topic area scores. In both cases the scores are collated to produce main criteria scores.
- Did not include behavioural questions by default, but rather the interviewer asked either a behavioural question or a situational question, depending on the relative scores on the two different SJE rating scales.

Details of these changes and rationale behind the developments are given in Appendix C.

3 The evaluation criteria

In order to determine whether the psychometric assessment methods shortlisted were suitable for future train driver selection, a framework was developed to evaluate the methods from a variety of perspectives, including validity, reliability and fairness. This framework is based on the standards specified in the European Federation of Psychologists Association (EFPA) review model for the description and evaluation of psychological tests (Bartram, 2008; Bartram, 2010). The EFPA Test Review Criteria were largely modelled on the form and content of the British Psychological Society's (BPS) test review criteria and criteria developed by the Committee of Test Affairs (COTAN) of the Dutch Association of Psychologists (NIP).

The benchmark for the judgement of fairness was based on the 4/5ths rule published by the US Equal Opportunity Employment Commission. This was used because it was the only objectively measurable criterion that was available to make a judgement against. However, it is important to be aware that it is not a legal definition of discrimination but rather a 'rule of thumb' that can be used to identify serious discrepancies. The Equal Opportunity Employment Commission is a federal agency of the United States Government. The 4/5ths rule is not a rule of UK or European law and it has not be adopted or recommended by any official body with jurisdiction in Great Britain.

The evaluation criteria are a structured way of considering all the merits and limitations of each assessment method. In order for a psychometric assessment method to be recommended it should ideally reach an acceptable level on each of the evaluation criteria outlined in the table below. However, the overall case needs to be considered and if an assessment method is suitable in most respects then it can still be recommended even if weaker in some other respects. The most important consideration is the ability of each assessment method to measure the job relevant attributes effectively (ie validity and reliability) because these attributes are considered to be necessary to be a safe train driver.

Table 21 - Evaluation criteria for reviewing psychometric assessment methods

Evaluation criteria	What is it?	Guidelines for acceptance
Criterion validity	There are different types of validity. In the case of criterion validity, scores on an assessment method are compared to relevant performance criteria such as performance in training or in a job to determine whether assessment method scores are related to the criteria. The stronger this relationship is then the stronger the evidence that the score can be used to predict subsequent performance. Criterion validity is usually reported using the correlation coefficient or validity coefficient which ranges from -1 to +1 depending on the intended direction of prediction: positive if higher scores are associated with higher performance, negative if lower scores are associated with higher performance.	Criterion validity should be in excess of 0.20. In addition, the validity coefficient should be as good as, or better than, the existing psychometric assessment method for the relevant criteria (if applicable).
Content validity	This approach to validity involves the evaluation of the assessment method by subject matter experts. Information is collected on whether the experts consider the assessment method to be an adequate representation of the psychological construct (eg attention) to be assessed, and whether the scores taken from the assessment method are an appropriate and accurate means of distinguishing between different levels of performance (eg ability to sustain attention over long periods of time).	The assessment method should be an adequate representation of the skill or ability to be assessed. Content validity is not a statistical concept and so cannot be described in terms of a correlation co-efficient.
Construct validity	Construct validity is about whether an assessment method measures what it claims to measure. Information collected from content and criterion validity studies contribute to evidence of construct validity by providing a better understanding of what an assessment method measures and what its limitations are. Construct validity has two subtypes; convergent validity and discriminant validity. Convergent validity is measured by the extent to which similar measures of a construct produce similar results. Discriminant validity is measured by the extent to which measures of different constructs do in fact measure different things (ie do not correlate significantly).	Evidence that the assessment method measures the psychological construct, eg through factor analysis, significant correlations with measures of the same constructs, and non-significant correlations with measures of different constructs.

Evaluation criteria	What is it?	Guidelines for acceptance
Face validity	Face validity refers to the acceptability of the method to the users, including candidates and assessors and what the user thinks the method is measuring. This is important in obtaining the candidate's co-operation and commitment. It is not a true measure of validity, as it means that the assessment method 'looks like' it will work, as opposed to 'has been shown to work'. However, it is important because it can affect candidate motivation and behaviour.	The assessment method should appear to be measuring the relevant psychological construct or selection criterion and should be acceptable to the candidates.
Reliability	Reliability is about how consistent and trustworthy a measurement is. If an assessment method is reliable, it will give the same measurement over time and on different occasions. There are several different types of reliability. Internal consistency evaluates the extent to which the individual elements within an assessment give consistent results. Internal consistency is usually expressed as a reliability coefficient, or Cronbach's alpha (α). Other types of reliability are parallel form, inter-rater and stability of test-retest. Note that estimates of these different types of reliability may not agree. For example, it is possible to find measures with good internal consistency and parallel form validity but poor stability over time.	All types of reliability are commonly expressed as correlation coefficients and should be at least 0.8 (ability measures) or 0.6 (behavioural measures).
Fairness	Does not unjustifiably discriminate between groups based on protected characteristics (eg age, gender, ethnicity, and disability). The US Equal Employment Opportunity Commission's 4/5ths rule has been used for over 20 years in applied psychology and employment law (Roth et al., 2006). The rule states that there is adverse impact when the protected group pass rate is less than 80% of the highest scoring group's pass rate.	The pass rate for a protected group should be at least 4/5^{ths} of the pass rate of the majority or advantaged group unless there is a difference that can be justified on job performance grounds.
Administration time	The time it takes to administer the selection method including practice examples.	The selection method should have an appropriate administration time to support its use. The whole assessment centre process should fit within one day.

Evaluation criteria	What is it?	Guidelines for acceptance
Costs	Start-up costs (licences, software, hardware, and training) and recurring costs (re-licensing, maintenance, and renewing materials).	The cost of the whole assessment centre process per candidate should not be unacceptably higher per head than the current process.

4 Evaluation study method

4.1 Participant characteristics

A large number of companies were approached to participate in the validation study through communication at the project steering group and via the Association of Train Operating Companies (ATOC). Of the companies approached, 17 agreed to release drivers and trainees to participate in the trial.

The aim was to include as many participants as practically possible during the trial period (September 2010 to May 2011), and to ensure good representation of females, older people and people from ethic minority groups. These people represent minorities within the train driver workforce in terms of the key 'protected characteristics' of gender, age, and ethnicity (Equality Act, 2010 (UK)). The trials were extended by six months to specifically target members of these groups so that it would be possible to assess the fairness of the methods. The breakdown of the sample collected is considered to be representative of the train driver population where females, older people and people from ethnic minority groups are under-represented.

Participants involved in the trials conducted during project T628 (RSSB, 2009) were also targeted in order to try and obtain a sample of participants who had attempted all of the assessment methods.

Participating drivers and trainee drivers were asked to fill in questionnaires on demographic information including gender, ethnicity, age, experience on the railway, role, education level, whether they have dyslexia and their first language.

A total of 146 participants (drivers, trainee drivers and failed candidates) took part in the trials. Demographic characteristics of the sample are presented below, with information relating to protected characteristics (gender, ethnicity, age) presented first. Not every participant trialled every method because the drivers and trainees could not always be released for long enough to sit all of the assessment methods and there were not enough interviewers to

conduct interviews with all participants. Demographic information relating to the sample for each method is provided in Appendix D.

The total sample included 126 males and 20 females.

Table 22 - Ethnicity of trial participants

Ethnicity	TOTAL	
	N	*%*
White British	125	86
White Irish	3	2
White other	2	1
Indian	4	3
Pakistani	2	1
Black Caribbean	7	5
Mixed race	3	2
Total	**146**	**100**

Table 23 – Age and experience demographic characteristics of the trial sample

	Complete trial sample				
	Mean	*Min*	*Max*	*SD*	*N*
Age	41.62	25	61	7.80	146
Years on railway	10.09	2	38	8.95	143

Table 24 – Role of trial participants

Role	TOTAL	
	N	*%*
Train drivers	109	75
Trainee drivers	16	11
Driver managers	4	3
Failed candidates	7	5
Other	8	6
No information provided	2	1
Total	**146**	**100**

Table 25 – Educational level of trial participants

	TOTAL	
Educational level	**N**	**%**
No formal qualifications	13	9
GCSEs	45	31
GNVQs	36	25
A-levels / Scottish highers	38	26
Degree	12	8
No information provided	2	1
Total	**146**	**100**

The total sample included four people with dyslexia (three people did not provide information on dyslexia), and six who had a first language other than English (one person did not provide information on first language).

Operational performance data were available for the majority of the sample and were collected using the data collection form and training records. In addition, training performance data and manager ratings of behaviour were collected.

All the above data were collected and stored securely in accordance with the Data Protection Act 1998 and the British Psychological Society's Ethical Code of Conduct (British Psychological Society, 2009).

RSSB produced trial information leaflets and consent forms to ensure that full informed consent was obtained from each trial participant.

As a gesture to help motivate participants to try their best during the trial, RSSB offered vouchers to the three participants with the highest scores on the trialled methods.

4.2 Sample size, power, and precision

The aim was to include as many participants as practically possible during the trial period, and to include at least seven participants belonging to each of the 'protected characteristics' groups so that a basic statistical comparison of test scores between groups could be conducted. A target overall sample size of 100 was used as recommended in guidance by the Commission for Racial Equality (Commission for Racial Equality, 1992).

The demographic characteristics of the trial sample were monitored throughout the trial, and the trialling period was extended by six months to target members of specific demographic groups.

Please see the limitations section below for further discussion of the sample size.

4.3 Measures and covariates

The measures in this research consisted of job performance data for participants and assessment method scores from each of the methods trialled and from the existing assessment methods.

Job performance measures consisted of operational driving performance data, training performance data and manager ratings of behaviour. The job performance measures collected represented a mix of subjective and objective measures. RSSB took all reasonable steps to ensure that the data provided was consistent and evidence-based by providing clear definitions of rating scale points and providing behavioural markers (observable indications of performance) for the manager ratings of behaviour.

Job performance data was consolidated where possible with job performance data collected in the previous trials (RSSB, 2010) in order to maximise the sample. For 'overall train handling, application of rules', 'speeding record' and 'attitude to work and people' the scales were the same in both trials. In this case performance data from the previous trials could be used without modification if there was no more recent performance data.

For 'SPAD record, 'collisions', 'station overruns', 'station disregards', 'operation of safety systems', preparation/disposal of trains' and 'workplace formal assessments' the scale was updated for the T948 trial to try and obtain greater sensitivity. In each case the two scales were compared and the job performance scores from the previous trials were mapped on to the new scales. If any drivers had job performance data from both the previous and recent trials then the recent data was used.

A complete list of all assessment method scores and job performance measures (including consolidated performance measures) is provided in Appendix E.

4.4 Data collection procedure

4.4.1 Assessment method data

Assessors (RSSB staff and assessment centre staff from the T948 steering group) were trained in advance of the trials in how to correctly and consistently administer and score (if necessary) each assessment method according to the procedure outlined in the relevant test manual. Assessors interviewing participants as part of the trial were required to have the BPS Level A qualification in psychometric testing, to attend training at RSSB, and to record a practice interview in advance of the trials for RSSB quality checking.

The assessment methods were administered according to a trial day plan to ensure consistency in the order in which methods were trialled. As two tests of vigilance were included, the trial plan was amended part way through the trials to swap the order of the vigilance tests to counterbalance any order effects. Two versions of the SJE and WCT were trialled. Participants were randomly assigned one version of these tests so that roughly half of the participants completed each version.

All methods were administered in conditions consistent with good practice in test administration.

Assessment method data for the Group Bourdon, DTG, TRP1, TRP2 and CBI, were also extracted from the RACF database to be used in the trial analysis.

4.4.2 Job performance data

Job performance data collection forms and associated guidance were distributed to the managers of drivers and trainees participating in the trial. Managers were requested to complete the forms once the participants had agreed to participate and signed the consent form.

For participants who had been involved in the previous trial (T628) it was possible to use some of the job performance data collected as part of those trials. More information on how performance data from the previous and current trials was combined is provided in Appendix E.

4.4.3 Assessor and participant feedback

A semi-structured discussion session was built into the trial day plan in order to collect feedback from all participants on each of the assessment methods (eg how relevant they are to the train driver role) at the end of each trial session. Each assessor was requested to complete a questionnaire following their involvement in the trials to provide their feedback on their experience of using each method, and their perception of its face validity.

4.5 Research design

This was a concurrent validation study which aimed to determine whether the assessment method scores were significantly related to job performance measures. Assessment method scores were collected from existing train drivers and trainee train drivers and job performance data were collected from their managers at the same time.

Expected correlations between assessment method scores and job performance measures were documented in advance of the trial (Appendix F outlines the predicted relationships).

4.6 Overview of data analysis method

Data were input, checked and consolidated (where possible) with data from the previous trial. Basic descriptive analysis was carried out. Qualitative analysis was conducted of the assessor and participant feedback collected during the trial.

The analysis looked at the performance of each assessment method against each of the evaluation criteria (see Section 3 of Annex 3). Different sources of information were consulted for different evaluation criteria.

4.6.1 Construct validity

To assess whether the assessment methods measured what they claim to measure a range of relevant information was considered.

To examine whether there was evidence of consistency between the theoretical construct(s) and the score(s) produced by the assessment methods, relevant information was sought from test manuals where available (ie the design of the test) and factor analyses.

Pearson product moment correlations (referred to as Pearson correlation throughout this report) were also calculated using trial data to examine the relationships between test scores collected in this trial, measures that should be theoretically related (convergent validity), and those that should not be related (divergent validity).

Correlations between manager ratings of behaviour and SJE and MMI measurements of the same behaviour were used to gain an indication of convergent validity in this trial. These correlations could also be considered an indication of criterion validity as the manager ratings were a form of performance criteria.

4.6.2 Content validity

Content validity was examined qualitatively to look at the degree of agreement between the content of the assessment method and the definition of the selection criterion that it was supposed to measure. Where relevant, the views of other authors were taken into account.

4.6.3 Criterion validity

Specific predictions were made in advance of the data analysis about which job performance measures should correlate significantly with each assessment method score and in what direction. All of these expected relationships are tabulated in Appendix F.

For all expected relationships, Pearson correlations were calculated between the assessment method score and the relevant job performance measure. In the presentation of results, only predicted relationships are presented. Any relationships that were not predicted have not been analysed. Predicted correlations that are statistically significant at the $p \leq 0.05$ level and in the expected

direction are shown in bold green font. These significant relationships are considered to be good evidence of criterion validity. Significant relationships in the unexpected direction are shown in bold red font. These relationships are considered to be evidence that the assessment method does not have good criterion validity.

4.6.4 Face validity

Face validity was assessed using feedback from assessors who administered the trial and trial participants.

4.6.5 Reliability

Where available, reliability was assessed using information published in the test manual. For the methods that were developed by RSSB (WCT, SJE and MMI) reliability was assessed using the Cronbach's alpha measure of internal consistency. In some cases, other measures were calculated to examine reliability in more detail or to assess other aspects of reliability, such as parallel forms. Where relevant, details of these analyses are provided in the findings section.

4.6.6 Fairness

Any information provided in the test manuals in relation to fairness was considered and analyses were run using the data collected in the trial. These analyses focussed on testing for average score differences between groups defined by the protected characteristics of gender, age and ethnic group. The following groups were compared:

- Males versus females
- White versus any other ethnicity
- Younger (50 years old and younger) versus older (51 years old and older)

The small numbers of participants in the minority groups meant that more sophisticated analysis with more differentiated groups was not possible.

The means and standard deviations of each test score were calculated for each group and compared statistically using t-tests or Mann Witney U if the data were not normally distributed. If there was a significant difference (at the $p \leq 0.05$ level) between the groups then it was considered to be evidence that different groups might perform differently on the assessment method. In this case, the acceptability of the assessment method depends on the justifiability of its use in terms of validity for assessment of a job relevant skills and the impact that it would have on candidate groups depending on where a cut-off is set. These issues were considered when designing the recommended assessment process and are discussed in Section 7 of Annex 3.

4.6.7 Administration time and cost

These factors were assessed initially when the assessment methods were shortlisted. In each case, the administration time and cost was considered to be acceptable. A further evaluation of administration time and cost of the whole train driver assessment process will be required as part of the impact assessment for the change if the recommendations of this work are taken forward to be implemented in RIS-3751-TOM.

4.7 Use of the data analysis findings for assessment method development

For the methods that were developed by RSSB, assessor feedback and the results of the data analysis contributed to decisions about improvements that could be made. The rationale for changes to the RSSB developed methods are documented in Appendices B and C.

For methods that were developed by RSSB, an iterative data analysis process was necessary. Where possible, criterion validity, reliability and fairness analyses were repeated each time the assessment methods were modified to determine whether the changes had had the required effect. In the validation study findings section, results are presented for the 'final' versions (or as close as possible, given subsequent recommended changes) of the psychometric assessment methods.

4.8 Limitations of the study

The design and method of this evaluation trial was the best compromise given the practical constraints of availability of participants, time and assessor resources.

The key limitations of the approach taken are:

- It was not possible to collect job performance data for every participant which reduced the sample size for the criterion validity analyses.
- Low numbers of participants in each of the 'protected characteristics' groups and of failed candidates limited the quality of the fairness analysis.
- Due to resource constraints, it was not possible for every participant to trial every assessment method. The sample sizes used to calculate test inter-correlations are therefore limited and there were few participants with scores for all of the different assessment methods trialled across the T948 and T628 projects.
- The variance in the job performance data was limited. The sample consisted largely of experienced drivers with adequate job performance. This made it difficult to set cut-off scores based on excluding those with poor performance.

5 Validation study findings

5.1 Introduction

These results are written in line with the American Psychological Association reporting guidelines.

Results are presented in a separate section for each assessment method trialled and are organised according to the evaluation criteria as detailed in Section 3.

The assessment methods developed by RSSB (WCT, SJE and MMI) have been refined on the basis of the results of the trial. Where possible, the results presented relate to the final version of each test that would be implemented if the recommendations are taken forward. This allows the reader to judge each assessment method as it would be if implemented.

As detailed in Section 4.1, not all participants completed all the assessment methods and not all had operational performance data. Consequently, the full sample was not available for all analyses. Appendix D contains the detailed demographic breakdown of the sample used for the analysis of each assessment method. This main findings section only contains high level information about the sample.

5.2 WAFV

These results relate to WAFV test form S2.

5.2.1 Sample

A total of 122 participants completed the WAFV. Detailed information on the demographic characteristics of the sample is provided in Appendix D. Out of the 122 participants who completed the WAFV, 80 had operational driving performance data and 34 had training performance data.

5.2.2 Construct validity

5.2.2.1 Convergent validity

The WAF test manual (Sturm, 2006) demonstrates the construct validity of WAFV through factor analysis with other tests of attention from the Vienna Test System battery including Cognitrone, Discrimination Test and Reaction test. The table below shows the factor structure of the WAF test battery (including the WAFV sub-test) obtained by principal component analysis with varimax rotation. For the sake of clarity loadings of less than 0.4 have been omitted.

Table 26 - Factor structure of the WAF test battery

	Factor 1	Factor 2	Factor 3
WAFA – Mean reaction time Subtest 1		.85	
WAFA – Mean reaction time Subtest 2	.43	.67	
WAFA – Mean reaction time Subtest 3		.70	
WAFA – Mean reaction time Subtest 4		.81	
WAFA – Mean reaction time Subtest 5		.78	.43
WAFA – Mean reaction time Subtest 6		.86	
WAFF – Mean reaction time Subtest 1	.58		
WAFF – Mean reaction time Subtest 2	.70		.47
WAFF – Mean reaction time Subtest 3	.83		
WAFS – Mean reaction time Subtest 1	.78		
WAFS – Mean reaction time Subtest 2	.79		
WAFS – Mean reaction time Subtest 3	.81		
WAFG – Mean reaction time Subtest 1	.91		
WAFG – Mean reaction time Subtest 2	.86		
WAFV – Missed reactions Test form 2			.71
WAFV – Missed reactions Test form 4			.82
WAFV – Missed reactions Test form 6			
WAFV – Missed reactions Test form 8			.52
COG – Mean time ‘correct rejection’	.72		
DT – Correct responses	-.67		
RT – Mean reaction time	.68		
RT – Mean motor time	.44		
SPM PLUS – Correct responses			

Sturm (2006) postulates that the content of the three factors can be clearly interpreted. Factor 1 represents the selectivity aspect of attention, while Factor 2 draws together tests that load primarily onto the short-term control of the intensity of attention (intrinsic and phasic alertness). Factor 3 comprises tests which require attention to be sustained over a lengthy period of time (vigilance). The WAFV scores load onto factor 3 which provides evidence that the assessment method measures vigilance, as opposed to other facets of attention.

For SPM Plus, which measures language-free general intelligence, there are no relevant loadings onto any of the 3 attention factors. From this, Sturm (2006) argues that the aspects of attention measured by the WAF test battery can be clearly distinguished from the 'G factor' of intelligence.

Additional evidence of construct validity is provided by Sturm (2006) in the WAFV test manual. The same data from the factor analytic approach was explored using a linear structural equation model which was drawn up on the basis of the theoretical model.

Table 27 - Linear structural equation model for the WAF tests extracted from WAF test manual

In the linear structural equation model for the WAF tests, the path weights are given as standardised regression coefficients. On the first level, the latent factors of Alertness (A), Vigilance (V), Focused Attention (F), Selective Attention (S) and Divided Attention (G) are estimated. At the second level, the latent factors of the intensity and selectivity aspects are estimated. The results of the modelling provide evidence that the empirical data fit the theoretically postulated model of vigilance. This evidence suggests that the method is a valid measure of vigilance (Sturm, 2006).

Pearson correlations were calculated to determine the strength of the correlations between WAFV and VIGIL which are both designed to assess vigilance. The correlations were only moderate ($r(121)$ = range from -.25 to .32, p <.001 to <.01) which suggests that these two vigilance tests are measuring only moderately related constructs.

5.2.2.2 Discriminant validity

Pearson correlations were calculated to determine the strength of the correlations between WAFV and other assessment methods which are designed to measure different constructs. WAFV scores significantly correlate with 2HAND scores ($r(121)$ = range from .21 to .37, p < .001 to .02) and with TAVTMB scores ($r(12)$ = range from .58 to .81, p < .001 to .05) although caution should be exercised when interpreting the latter as the sample only contains 12 participants. The correlations with 2HAND and TAVTMB may exist as these tests also involve making motor responses to visual stimuli presented on screen.

There were no significant correlations between WAFV and Group Bourdon, TEA-Occ, DTG, TRP1 or TRP2. This demonstrates that WAFV is measuring different constructs from tests which are not designed to measure vigilance.

Pearson correlations were also calculated to determine the strength of correlations *between* WAFV test scores to determine if all scores provide unique information within the psychometric assessment process. The results are provided in the table below.

Table 28 - Pearson correlations between WAFV scores

		Number of missed reactions	*Mean reaction time*	*Number of false alarms*
Number of missed reactions	*r*	1		
	P			
	N			
Mean reaction time	*r*	.59	1	
	P	<.001		
	N	122		
Number of false alarms	*r*	.44	.29	1
	P	<.001	.001	
	N	122	122	

Correlations between WAFV assessment method scores $r(122)$ range from .29 to .59, p= <.001.

These results suggest that the three test scores are measuring different but related constructs, and that there is value in using the different scores.

5.2.3 Content validity

Vigilance is defined as the ability to attend and respond to stimuli which occur relatively infrequently and over extended periods of time. This is demonstrated in WAFV as the stimuli occur very infrequently (5% of stimulus presentations) and the test duration is 35 minutes.

It has been argued strongly that the currently used tests of attention (the Group Bourdon and the DTG) do not give valid assessments of sustained attention or vigilance. In particular, several research papers have reported that short tests of focused attention do not predict vigilant performance (Robertson & Garavan, 2004; RSSB, 2005). The WAFV is considered to have stronger content validity than the possible alternative measures.

5.2.4 Criterion validity

Appendix Section G.1 provides details of all the correlations between WAFV test scores and performance data. Of the 15 relationships predicted between WAFV test scores and operational performance data, six reached significance, which is statistically more than would be expected by chance (only one would be expected by chance).

Notable results include:

- Number of missed reactions significantly correlates with train handling ($r(80) = -.21$, $p = .03$) and SPAD record ($r(80) = -.19$, $p = .05$).
- Mean reaction time significantly correlates with SPAD record *(r(80)= -.19, p = .05).*
- Number of false alarms significantly correlates with train handling ($r(80) = -.36$, $p = <001$), SPAD record ($r(80) = -.20$, $p = .03$) and speeding ($r(80) = -.30$, $p = <.001$).

The evidence suggests that WAFV has good criterion validity as the WAFV scores correlated with the relevant operational performance ratings as predicted.

It is noteworthy that WAFV is embedded in modern attention theory and is commonly used by neuropsychologists for assessing differences between healthy and brain damaged people. As such, this may be the first traffic validation study to be undertaken with this test which provides criterion validity evidence for driving applications.

5.2.5 Face validity

Qualitative comments collected during the trials from participants and assessors demonstrate clear support for the face validity of the WAFV. A key theme arising from the comments was that it was clear to see how the WAFV measured constructs that are relevant to the train driver role. Specifically, participants stated that they believed the test drew on the skills needed to successfully notice changing signal aspects encountered during train driving.

5.2.6 Reliability

The WAFV test manual reports a Cronbach's alpha of 0.96 for test form S2 vigilance visual long form (Sturm, 2006). Further, the inter-correlations between the first and second half of the test (split-half reliability) confirm the reliability of this test:

- Missed reactions r = .826
- False alarms r =.499
- Mean reaction time r =.838

This evidence demonstrates that WAFV has excellent reliability.

5.2.7 Fairness

Table 29 shows the means for the main WAFV scores according to gender, age and ethnicity.

In terms of gender effects, no significant differences were found between males and females.

In terms of differences between ethnic groups, no significant differences were found between whites and other ethnic groups.

There was no significant difference according to age group on any of the scores of interest.

Table 29 – WAFV scores according to gender, ethnicity and age

			WAFV scores					
			Number of missed reactions		Mean reaction time		Number of false alarms	
Protected characteristics		*N*	*Mean score*	*SD*	*Mean score*	*SD*	*Mean score*	*SD*
Gender	*Males*	111	.95	2.71	392.50	82.70	4.83	11.90
	Females	11	.09	.30	376.91	65.37	2.82	6.15
Ethnicity	*White*	113	.88	2.69	391.38	83.10	4.27	11.36
	Other	9	.89	.78	387.56	55.58	9.33	12.86
Age	*Up to 50 years*	107	.79	2.12	385.87	77.37	4.16	10.15
	Over 51 years	15	1.53	4.87	428.40	99.70	8.13	4.82
Overall sample		122	.88	2.59	391.10	81.18	4.65	11.50

5.2.8 Administration time and cost

The WAFV, being a computerised assessment method, provides a very economical method of administration and scoring. The administrator's time is saved because the instructions at the beginning of the test are standardised and raw and norm values are calculated automatically (Sturm, 2006).

The time required to complete WAFV test form S2 is 35 minutes with practice examples. The test administrators during the trial reported that WAFV was straightforward to administer.

5.2.9 Overall conclusion about the WAFV

WAFV test form S2 meets a good standard on all of the evaluation criteria. There were a notable number of meaningful significant correlations with predicted operational driving performance measures including overall train handling, SPAD records and speeding. Significant correlations between WAFV scores and behaviour were found as expected, and each either met, or exceeded the r = .2 level specified in the EFPA evaluation guidance.

There was strong evidence that the WAFV has good content and construct validity. In a factor analysis with other tests of attention, WAFV loaded onto a different factor than tests of selective and divided attention. Further, the linear structural equation model demonstrated that the WAFV sub-test of the WAF test battery correlated with latent factors of vigilance, as opposed to other facets of attention.

There was excellent evidence of reliability, for both internal consistency and split-half reliability, and the analysis of fairness did not highlight any major concerns. In addition, there was high acceptance of this test by train drivers as they felt that it was an appropriate measure of vigilance.

WAFV demonstrated stronger construct validity and more consistent links with operational performance records than VIGIL. It is therefore recommended that WAFV should be used to assess vigilance in the future train driver psychometric assessment process.

5.3 VIGIL

These results relate to VIGIL test form S2 Müggenburg 33.

5.3.1 Sample

A total of 123 participants completed VIGIL. Detailed information on the demographic characteristics of the sample is provided in Appendix D. Out of the 123 participants who completed VIGIL, 81 had operational driving performance data and 34 had training performance data.

5.3.2 Construct validity

The VIGIL test manual (Puhr, 2004) reports the examination of various attention tests and others by Sturm & Büssing (1990) performed with patients with cerebral damage in the left hemisphere versus the right hemisphere. These results showed a clear correlation between the attention levels obtained in the VIGIL test form S2-Müggenburg 33 and other attention-related capacities of patients with damages in the right hemisphere of the cerebrum. These findings demonstrate that VIGIL is capable of identifying candidates who have unacceptable levels of attention for safety critical roles.

5.3.2.1 Convergent validity

Pearson correlations were calculated to determine the strength of the correlations between VIGIL and WAFV which are both designed to assess vigilance. Surprisingly, the correlations were moderate ($r(121) = -.25$ to $.32$, $p < .001$ to $< .01$) which suggests that these two vigilance tests are measuring only moderately related constructs.

5.3.2.2 Discriminant validity

Pearson correlations were calculated to determine the strength of the correlations between VIGIL and other assessment methods which are designed to measure different constructs. VIGIL significantly correlated with 2HAND ($r(121) = -.39$ to $-.25$, $p < .001$ to .005), with TAVTMB ($r(12) = -.61$, $p < .03$) and TRP2 ($r(55) = .29$, $p = .03$). Like WAFV, VIGIL also correlated with 2HAND and TAVTMB which may be due to the nature of the task, where candidates are required to make a motor response (pressing a button) to visual stimuli.

There were no significant correlations between VIGIL and Group Bourdon, TEA-Occ, DTG or TRP1. This demonstrates that VIGIL is measuring different constructs from tests which are not designed to measure vigilance.

Pearson correlations were also calculated to determine the strength of correlations between VIGIL test scores to determine the extent the individual scores provide unique information within the psychometric assessment process. The results are provided in the table below.

Table 30 - Pearson correlations between VIGIL scores

		Number of correct	*Number of incorrect*	*Mean value of time correct*
Number of correct	*r*	1		
	P			
	N			
Number of incorrect	*r*	-.06	1	
	P	.50		
	N	123		
Mean value of time correct	*r*	-.22	-.30	1
	P	.01	.001	
	N	122	122	

Correlations between VIGIL assessment method scores $r(123)$ = range from -.06 to -.30, p= <.001 to .50.

These results suggest that the three test scores are measuring different but related constructs, and that there would be value in using the different scores.

5.3.3 Content validity

Content validity is demonstrated in VIGIL as the stimuli occur infrequently compared to other test forms and the test duration is 35 minutes.

The VIGIL test manual (Puhr, 2004) states that tasks aimed at measuring the ability to 'maintain watchfulness in stimulus-deficient situations' should be independent of a person's intelligence. Puhr (2004) considers VIGIL to meet this criterion due to the simple differentiation required in the test.

Moreover, the VIGIL test manual (Puhr, 2004) claims that interpretation of performance during the course of the test indicates a changing performance of the respondent in vigilance tasks. Deteriorations in performance are characterised by an increased reaction time and a decreased number of correct reactions to crucial stimuli, which are depicted in the ascent or descent of the straight lines of regression.

5.3.4 Criterion validity

Appendix section G.2 provides details of all the correlations between VIGIL test scores and performance data. Of the 15 relationships predicted between VIGIL test scores and operational performance data, none reached significance.

In summary, the evidence of criterion validity from the T948 trials is generally poor and does not satisfy the EFPA evaluation guidance. The VIGIL test manual (Puhr, 2004) does not provide information on the criterion validity of this assessment method with relevant job performance measures so the results of T948 must be considered alone.

5.3.5 Face validity

Qualitative comments collected during the trials from participants and assessors were varied for the face validity of VIGIL. Whilst some participants reported that VIGIL was credible as a measure of vigilance, other participants felt that the test had limited relevance to the role of a train driver. Further, a small number of the participants reported to experience discomfort and eye strain during the assessment.

5.3.6 Reliability

VIGIL has very good levels of internal consistency. The VIGIL test manual reports the following ranges for Cronbach's Alpha for test form S2:

- Number of correct: r = 0.65 – 0.82
- Number of incorrect: r = 0.69 – 0.77
- Mean reaction time: r = 0.87 – 0.89

This evidence demonstrates that VIGIL provides reliable measures of vigilance.

5.3.7 Fairness

There were no significant differences in scores according to gender, ethnic group or age in the T948 trial data. Table 31 shows the mean scores for each group.

Table 31 – Mean VIGIL scores according to gender, ethnicity and age

			VIGIL scores					
			Number of correct		Number of incorrect		Mean value of reaction time correct	
Protected characteristics		*N*	*Mean score*	*SD*	*Mean score*	*SD*	*Mean score*	*SD*
Gender	*Males*	112	28.66	3.37	24.80	105.27	.58	.10
	Females	11	30.00	1.95	12.27	14.04	.58	.09
Ethnicity	*White*	114	28.86	3.32	22.70	101.95	.58	.10
	Other	9	27.78	2.91	36.00	85.01	.55	.07
Age	*Up to 50 years*	108	28.65	3.42	25.64	107.17	.58	.09
	Over 51 years	15	29.73	1.94	9.60	10.08	.58	.19
Overall sample		123	28.44	4.87	23.68	11.56	.58	.14

Further, the VIGIL test manual (Puhr, 2004) states that there is nothing which implies the assessment method is unfair or discriminates against certain persons. Puhr (2004) also suggests that it is unlikely that persons inexperienced with computers will be put at a disadvantage as the control panel is basic and candidates are only required to press one button.

5.3.8 Administration time and cost

The time required to complete VIGIL test form S2 is 35 minutes with practice examples. The test administrators during the trial reported that VIGIL was straightforward to administer.

5.3.9 Overall conclusions about VIGIL

Evidence of the criterion validity of VIGIL test form S2 was generally poor for the assessment of train drivers. None of the main test variables correlated with train driving performance.

There was evidence that the test is reliable, fair and has good content and construct validities. There was inconsistent acceptance of this test by participants in the trial and the majority felt that WAFV was more relevant to the train driving role.

In summary, the positive evidence for WAFV outweighed the positive evidence for VIGIL. Therefore, VIGIL is not recommended as a measure of vigilance in the future driver psychometric assessment process.

5.4 2HAND

These results relate to 2HAND test form S1 (ten runs with joysticks).

5.4.1 Sample

A total of 123 participants completed 2HAND. Detailed information on the demographic characteristics of the sample is provided in Appendix D. Operational driving performance data were available for 81 participants and 34 had training performance data.

5.4.2 Construct validity

5.4.2.1 Convergent validity

2HAND is the only assessment method which was trialled as part of this research to measure hand coordination. Therefore, it is not possible to use the trial data to look at correlations with other assessment methods which measure similar constructs. The 2HAND test manual (Puhr, 2003) does not provide information on convergent validity.

5.4.2.2 Discriminant validity

Pearson correlations were calculated to determine the strength of the correlations between 2HAND and other assessment methods which are designed to measure different constructs. 2HAND significantly correlates with WAFV ($r(121)$ = range from .21 to .37, $p < .001$ to .02) and VIGIL ($r(121)$ = range from -.39 to -.25, $p < .001$ to .005). This might be because the 2HAND task requires attention as well as hand coordination, though it is not the main purpose of the test. Further, these tests all require the candidate to make a motor response to visual stimuli. However, the correlations are moderate.

It is not possible to present the intercorrelations with TEA-Occ or TAVTMB as there were too few participants with scores for both assessment methods (N <10). There were no significant correlations between 2HAND and Group Bourdon, DTG, TRP1 or TRP2.

Considering this evidence, 2HAND is the only test which provides a measure of hand coordination though it may also assess attention.

Pearson correlations were calculated to determine the strength of correlations between 2HAND test scores. The results are provided in Table 32 below.

Table 32 - Pearson correlations between 2HAND scores

		Total mean duration	*Total mean error duration*	*Total percent error duration*	*Coordination difficulty*
Total mean duration	*r*	1			
	P				
	N				
Total mean error duration	*r*	.06	1		
	P	.52			
	N	123			
Total percent error duration	*r*	-.40	.79	1	
	P	<.001	<.001		
	N	123	123		
Coordination difficulty	*r*	.43	-.002	-.19	1
	P	<.001	.99	.03	
	N	123	123	123	

Correlations between 2HAND assessment method scores $r(123)$ = range from .-.40 to .79, p= <.001 to .99.

These results suggest that two of the three test scores are measuring different but related constructs, and that there could be value in using the different scores. However, the correlation between total percent error duration and total mean error duration is very high ($r(123) = .79$, p= <.001) so there would only be value in recommending one of these scores.

5.4.3 Content validity

The 2HAND focuses on two components of ability: sensory motor eye-hand coordination and coordination between the left and right hand. Eye-hand coordination is an important skill for smooth train handling. The difficulty in coordinating both hands arises from the need to make a correct visual assessment of the proportion of left- and right-hand controlled deviation from the target and to make adjustments accordingly. This aspect of coordination may be less important to the train driving role.

5.4.4 Criterion validity

The 2HAND table in Appendix Section G.3 provides details of all the correlations between 2HAND test scores and performance data. Of the 12 relationships predicted between 2HAND test scores and operational performance, two reached significance, which is

statistically more than would be expected by chance. Less than one significant correlation would be expected by chance.

Notable results include:

- Total mean duration significantly correlates with train handling ($r(81) = -.28$, $p = .03$).

In summary, the evidence suggests that 2HAND has reasonable criterion validity for the total mean duration score but the other three test scores did not significantly correlate with relevant performance measures as predicted.

In the 2HAND test manual, Puhr (2003) reports that Karner and Neuwirth (2000) evaluated the criterion validity of 2HAND within a traffic psychology study. The authors demonstrated a significant correlation between the overall mean duration score and the global assessment of the driving-competence in driving samples of 0.497 (p= <0.01). Furthermore respondents with a percentile rank of < 33 achieved a lower assessment of their driving-competence in driving samples compared to other respondents (Mann-Whitney U-test: z = 2.224, p <0.05). The test therefore appears to be a valuable tool in traffic psychology assessment.

5.4.5 Face validity

Qualitative comments collected during the trials from participants and assessors demonstrate mixed support for the face validity of the 2HAND. Some participants felt that the two-handed control and diagonal section of the track were not relevant for testing a person's ability to drive a train. There was also some dissatisfaction expressed with the delicate controls.

5.4.6 Reliability

There is excellent evidence of the reliability of 2HAND. The internal consistency (Cronbach's Alpha) of test form S2 ranges between 0.93 and 0.97 (Puhr, 2003).

5.4.7 Fairness

Assuming respondents do not have severe motor deficits, which would impair controlled coordination, it should not be expected that particular respondents would be systematically discriminated against (Puhr, 2003).

Table 33 shows the mean test scores that were observed according to gender, age and ethnicity groups.

There were no significant differences based on age for the relevant test variables.

Non-white ethnic groups took significantly longer to complete the test than white participants on the overall mean duration test score {t(121) = -2.267, p = 0.03}.

Females had a significantly higher mean error duration (a worse score) than males {t(121) = -2.272, p = .03}.

There were no other significant differences according to ethnicity or gender.

Table 33 – Mean 2HAND scores according to gender, ethnicity and age

			2HAND scores							
			Total mean duration		Total mean error duration		Total percent error duration		Coordination difficulty	
Protected characteristics		*N*	*Mean score*	*SD*	*Mean score*	*SD*	*Mean score*	*SD*	*Mean score*	*SD*
Gender	*Males*	112	28.18	9.78	.91	.63	3.74	3.29	3.05	.98
	Females	11	31.84	5.37	1.41	1.17	4.30	3.15	3.38	.93
Ethnicity	*White British*	114	27.97	9.26	.96	.71	3.86	3.36	3.05	.98
	Other	9	35.31	10.61	.98	.57	2.89	1.37	3.33	.91
Age	*Up to 50 years*	108	28.09	9.63	.94	.70	3.82	3.38	3.12	1.00
	Over 51 years	15	31.52	8.30	1.10	.75	3.59	2.31	2.77	.73
Overall sample		123	28.51	9.52	.96	.70	3.79	3.26	3.08	.98

5.4.8 Administration time and cost

The administration time of the 2HAND is ten minutes although a practice phase adds extra time.

5.4.9 Overall conclusions about the 2HAND

There is moderate evidence of criterion validity for the 2HAND. The total mean duration score correlates significantly with predicted operational driving performance including train handling and train handling during abnormal situations.

However, there were no significant correlations between total mean error duration, total percent error duration and coordination difficulty and relevant job performance measures. However, total percent error duration is a control measure against cheating.

This assessment method has good reliability and incremental validity within test scores. However, train drivers in the trials did not

think this test was typical of the coordination required to drive a train.

As the TDLCR (2010) requires that hand coordination be assessed, a measure of hand coordination must be included in the future psychometric assessment of train drivers. The 2HAND is currently used to assess hand coordination in train driver selection in several other European Member States as well as Queensland National in Australia. Therefore, it is recommended that 2HAND is included in the new psychometric assessment process. However, it should only be used to exclude a small number of candidates who have extremely poor hand-coordination.

5.5 WCT

Two versions of the WCT were specially designed to measure the ability to communicate effectively in writing: version 1 and version 2.

Due to the nature of the changes made to the original WCT, it was not possible re-run the analysis on the final version. Therefore, the results presented in this section refer to the original versions of the WCT that were trialled.

5.5.1 Sample

A total of 142 participants completed the WCT; 69 completed version 1 (61/69 had relevant primary performance data and 23/69 relevant secondary performance data) and 73 completed version 2 (62/73 had relevant primary performance data and 18/73 relevant secondary performance data).

5.5.2 Construct validity

The WCT was designed to assess four aspects of written communication skill: accuracy, written comprehension, legibility and structure (formed of logical sequencing and concise). For further information about the development of the WCT see Annex 3, Appendix B.

5.5.2.1 Convergent validity

Pearson's correlations were calculated to determine the strength of correlation between the overall written communication score and the written communications rating given by managers. Appendix section G.8 provides full details of these correlations.

Version 1

Of the 16 relationships predicted between the WCT and managers' ratings of written communication, three reached significance. Only one correlation would be expected by chance. Two relationships were significantly negatively correlated.

The notable results are as follows:

- The overall WCT score correlates significantly with the overall written performance rating given by managers ($r(59) = .22$, $p = .04$).
- The overall WCT score correlates significantly with the legibility score given by managers on written communication ($r(59) = .39$, $p = <.01$).
- The WCT score for the details section correlates significantly with the total performance rating given by managers on written communication ($r(59) = .37$, $p = <.01$).

Version 2

Of the 16 relationships predicted between the WCT and managers' ratings of written communication, seven reached significance. Only one would have been expected by chance. Two were significantly negatively correlated.

The notable results include:

- The overall WCT score correlates significantly with the total performance rating given by managers on written communication ($r(60) = .27$, $p = <.05$).
- The WCT score for concise structure correlates significantly with the conciseness performance rating given by managers ($r(60) = .33$, $p = <.01$).
- The overall WCT score correlates significantly with the logical performance rating given by managers ($r(60) = .31$, $p = <.01$).

The results demonstrate that the variables of both WCT versions have mostly positive, albeit not all significant, correlations with overall performance ratings given by managers on the written communications performance factor.

In conclusion, both WCT versions have good convergent validity as there is evidence that the variables correlated with the primary performance ratings as predicted.

5.5.2.2 Discriminant validity

Pearson's correlations were calculated to determine the strength of correlation between the four aspects of written communication. Table 34 and Table 35 provide details of these correlations:

Table 34 – Pearson correlations between WCT version 1 scores

WCT version 1		WCT total scores			
WCT total scores		*WCT: Detail section*	*WCT: Summary section*	*WCT: Legibility section*	*WCT: structure section*
WCT: Detail section	*r*	1			
	p				
	N				
WCT: Summary section	*r*	-.03	1		
	p	.79			
	N	69			
WCT: Legibility section	*r*	-.03	-.17	1	
	p	.79	.18		
	N	69	69		
WCT: Structure section	*r*	.07	-.04	-.05	1
	p	.58	.78	.67	
	N	69	69	69	

Table 35 - Pearson correlations between WCT version 2 scores

WCT version 2		WCT total scores			
WCT total scores		*WCT: Detail section*	*WCT: Summary section*	*WCT: Legibility section*	*WCT: structure section*
WCT: Detail section	*r*	1			
	p				
	N				
WCT: Summary section	*r*	.05	1		
	p	.70			
	N	69			
WCT: Legibility section	*r*	NP	NP	NP	NP
	p				
	N	69	69		
WCT: Structure section	*r*	-.08	.20	NP	1
	p	.52	.09		
	N	69	69	69	

These results show one positive but not significant correlation and several negative correlations between the sections of WCT version 1. The results for version 2 show two positive correlations, one negative correlation and a few correlations that were not able to be calculated due to no score variance.

The results suggest a mixed picture for the WCT. Overall, each of the four WCT components on both versions of the test measure different constructs of written communication; however there is likely to be overlap in the skills being measured as they are subsets of a small aspect of written communication. These results also go some way to explain the relatively weak reliability of this test, (as described in Section 5.5.6 below).

5.5.3 Content validity

The WCT was specifically developed to measure written communication in response to concerns raised by the industry steering group about a lack of a standardised and structured way of measuring written communication.

The WCT was designed to provide a valid measure of written communication through coverage of four areas components of written communication (accuracy, written comprehension, legibility and structure).

- The details section is about accuracy of details being transferred onto the incident report form.
- The summary section is about comprehension – understanding what happened about being able to communicate that effectively in writing.
- Legibility is about having clear handwriting (a prerequisite to all other sections).
- Structure is about logical sequencing of events that occurred and also of how relevant the information is.

These four skills were considered the most important aspects of written communication and agreed by a group of SMEs prior to the trials.

As a train driver, writing ability is most aptly and critically demonstrated when filling out report forms. The WCT simulates a report form writing task that is based on a train driver incident report form so has clear content validity.

5.5.4 Criterion validity

5.5.4.1 Version 1

Of the three relationships predicted between the WCT and performance data, all three were positive and two reached significance. This is more than would be expected by chance.

The notable results are:

- The overall score on the WCT correlates with the job performance measure of trainee rules assessment ($r(21) = .52$, $p = <.01$)
- The overall score on the WCT correlates with the job performance measure of trainee traction theoretical ($r(21) = .49$, $p = <.01$)

5.5.4.2 Version 2

Of the three relationships predicted between the WCT and performance data two were positive but not significant and one was in the wrong direction. Version 2 did not correlate significantly or in the right direction with these job performance measures.

The variables of both versions of the WCT have mostly positive, albeit not significant, correlations with the performance factors. However, with small sample sizes for all of the job performance data with the WCT (version 1 n = 23, version 2 n = 18), the results of cannot be relied upon to make firm conclusions and it is more important to note the outcome of the relationships with managers ratings of communication with the WCT (described under construct validity, see Section 5.5.2.1 above).

5.5.5 Face validity

The qualitative comments collected from participants and assessors demonstrate clear support of using the WCT to measure written communication skills. The key theme was that the test appeared relevant to the role of a train driver and that the skills that needed to be demonstrated in the test were relevant for report writing. A few comments about the lack of clarity on a few pictures of the storyboards have informed and been incorporated into the development of the revised versions of the WCT.

5.5.6 Reliability

The number of items making up each section and the total WCT scale for each version of the WCT, and the corresponding alpha values, are provided in Table 36.

Table 36 - Items and alpha values for each WCT section

Test measure	Items	Version 1: Cronbach's Alpha	Version 2: Cronbach's Alpha
Overall total	16	.13	.43
Accuracy	4	.79	NP
Comprehension	9	.11	.47
Legibility	1	NA	NA
Structure	2	-.08	-.12

For both versions of the WCT, with all 16 items of the WCT scale included in the analysis, the results show that the alpha level did not reach the .7 level required, as specified in the evaluation guidance. The accuracy section on version 1 was reliable (4 items, α = .79).

Cronbach's alpha could not be computed for the legibility section as a minimum of two items are required for alpha levels to be computed.

For both WCT versions, the negative alpha for the structure section indicates that the two items within the structure scale (logical sequencing and conciseness) do not form a useful section because they are not measuring the same attribute.

It is important to note that another form of reliability (inter-rater reliability) was also calculated for this test. Three assessors separately rated eight of the same candidate report forms for each version of the WCT. These scores were then correlated to measure the level of consistency between each version. Table 37 below shows these results.

Table 37 - Inter-rater reliability of the WCT based on overall scores

			Assessor 1	Assessor 2	Assessor 3
WCT version 1	*Assessor 1*	*r*	1		
		p			
		N			
	Assessor 2	*r*	.97	1	
		p	.00		
		N	8		
	Assessor 3	*r*	.95	.98	1
		p	.00	.01	
		N	8	8	
WCT version 2	*Assessor 1*	*r*	1		
		p			
		N			
	Assessor 2	*r*	.76	1	
		p	.02		
		N	8		
	Assessor 3	*r*	.91	.90	1
		p	.00	.00	
		N	8	8	

The table above demonstrates that Pearson correlations for inter-rater reliability are very high (all below <.05 level, 2-tailed) for both versions of the WCT and between all the assessors.

The reliability analysis showed that both versions did not meet the expected Cronbach's alpha levels of internal consistency although inter-rater reliability was strong. The WCT is a short test, made up of four sections which have a total of only 18 items. This factor has greatly impacted reliability (alpha) levels. The in-depth analysis indicated that the cause of the low alpha levels was a lack of score variance and a few items being too easy (see Appendix B.2.4) and that most of the score variance related to the comprehension section.

Parallel forms reliability was also calculated using consolidated performance data on communication as a matching condition and also using performance ratings given by managers as a matching condition to pair the two samples of participants.

Fifty-five consolidated communications scores were available for those took version 1 and fifty-nine who undertook version 2. Candidates who undertook version 1 of the WCT were matched with those who undertook version 2 of the WCT based on similar consolidated performance criteria scores. This created 38 pairings. A paired-samples t-test was conducted to compare the scores on WCT version 1 and version 2. There was a significant difference in the scores for WCT v1 (M = 14.92, SD = 1.57) and WCT v2 (M = 15.71, SD = 1.58) conditions; t(37)= -2.18, p = .036 (2-tailed).

Sixty-two aggregated performance manager ratings on written communication were available for those took version 1 and sixty-one for those who undertook version 2. Candidates who undertook version 1 of the WCT were matched with those who undertook version 2 of the WCT based on these ratings. This created 40 pairings. A paired-samples t-test was conducted to compare the scores on WCT version 1 and version 2. There was a significant difference in the scores for WCT v1 (M = 14.58, SD = 1.75) and WCT v2 (M = 15.43, SD = 1.53) conditions; t(39) = -2.59, p = .013 (2-talied).

These results suggest that there is a slight difference between the two versions of the WCT. However, the standard deviations are quite small and there are no significant correlations between scores on both matched pairings. Also, with such low sample sizes, some items being too easy and the overall reliability of the WCT affected by the lack of score variance, it is possible that these differences may not be seen once the revised versions of the WCT have been implemented.

Fairness

Table 38 shows the descriptive statistics for the WCT version 1 scores by protected characteristics groups.

Table 38 – Mean WCT version 1 scores according to gender, ethnicity and age

WCT version 1			**WCT score**									
			Details section		**Summary section**		**Legibility section**		**Structure section**		**Overall total**	
Protected characteristics		*N*	*Mean Score*	*SD*	*Mean score*	*SD*	*Mean score*	*SD*	*Mean score*	*SD*	*Mean score*	*SD*
Gender	*Males*	59	3.81	.68	6.15	1.38	1.95	.22	2.93	.25	14.86	1.54
	Females	10	3.90	.32	5.60	1.27	2.00	.000	2.90	.32	14.40	1.35
Ethnicity	*White*	62	3.89	.45	6.15	1.41	1.95	.22	2.92	.28	14.92	1.49
	Other	7	3.29	1.50	5.43	.54	2.00	.00	3.00	.00	13.71	1.38
Age	*Up to 50 years*	60	3.90	.44	5.95	1.35	1.97	.181	2.95	.22	14.78	1.47
	Over 51 years	9	3.33	1.32	6.89	1.27	1.89	.33	2.78	.44	14.89	1.90
Overall sample		69	3.83	.64	6.07	1.37	1.96	.21	2.93	.26	14.80	1.51

Table 39 below shows the descriptive statistics for the WCT version 2 scores by protected characteristics groups.

Table 39 - Mean WCT version 2 scores according to gender, ethnicity and age

WCT version 2			WCT score									
			Details section		Summary section		Legibility section		Structure section		Overall total	
Protected characteristics		*N*	*Mean Score*	*SD*	*Mean score*	*SD*	*Mean score*	*SD*	*Mean score*	*SD*	*Mean score*	*SD*
Gender	*Males*	65	3.97	.17	6.92	1.51	1.91	.29	2.89	.31	15.69	1.61
	Females	8	3.88	.35	7.39	1.77	2.00	.00	2.88	.35	16.13	1.89
Ethnicity	*White*	65	3.95	.21	6.94	1.53	1.92	.27	2.89	.31	15.71	1.62
	Other	8	4.00	.00	7.25	1.67	1.88	.35	2.88	.35	16.00	1.85
Age	*Up to 50 years*	61	3.95	.22	7.13	1.48	1.92	.28	2.89	.32	15.89	1.58
	Over 51 years	12	4.00	.00	6.17	1.64	1.92	.29	2.92	.29	15.00	1.76
Overall sample		73	3.96	.20	6.97	1.54	1.92	.28	2.89	.32	15.74	1.63

There were two significant differences between groups on version 1 of the WCT.

- There was a significant difference on the overall score for version 1 (Mann Witney $U(2)= 114.50$, $Z = 142.50$, $p = .04$), where white participants scored slightly better than non-whites on the overall score for version 1.
- The older group of candidates scored lower on the detail section of version 1 ($U(2)=197.00$, $Z = -2.48$, $p = .01$).

5.5.7 Administration time and cost

The test takes ten minutes to complete; two minutes are provided for studying the picture storyboard and a further eight minutes for filling in the form. An additional two minutes are needed for assessors to read through the instructions with the candidates. It takes under five minutes to score each candidate report form (following training and practice of scoring) so in total the WCT comes within 15 minutes for administration to a group plus 5 minutes to score per candidate. Other than practical considerations, there are no restrictions to the number of candidates that can sit the WCT at any one time.

The cost of using the WCT and associated training costs are yet to be determined.

5.5.8 Overall conclusion about the WCT

There were strong relationships between WCT scores and manager ratings of written communication. These were the primary source of evidence for the WCT and both versions had good correlations as expected and exceeded the *r*= .2 level specified in the evaluation guidance.

This job performance data that was used to assess the criterion validity of the WCT were of secondary importance because the job performance measures are less directly relevant to written communication. There were moderate relationships between the WCT scores and measures of training performance. In spite of the positive results, there were small sample sizes for these analyses so the results cannot be relied upon to make firm conclusions.

The intercorrelations between the various WCT scores are low. The results suggest that each of the four WCT components on both versions of the test measure different constructs of written communication, which is as expected.

Reliability (internal consistency) for both versions of the WCT is below expected levels. However, by increasing the test length and making a few other amendments it is expected that the alpha of the comprehension section and of the overall test will increase. Whilst the WCT is weak on one aspect of reliability (Cronbach's alpha), it has very strong inter-rater reliability and there is evidence for its validity. So although the psychometric properties for the WCT are not strong overall, it is fit for the purpose it was designed for which is to assess written communication in a simple but more structured way.

The current method of assessing written communication is based on a form that candidates complete describing experiences on various topic areas that are then explored during the CBI. The industry group felt this form did not provide a very structured way in which to assess these skills. The WCT was developed to assess written communication in a more structured and systematic way and it does this well; however it is not working as a proper psychometric test because it assesses too many different aspects of writing skill with too few items. Written communication is not a safety-critical selection criterion because driver writing tasks are not part of the time critical driving task. However, there is still value in assessing written communication using the WCT for the purpose it is intended, particularly as the amendments to the scoring and pictures will go some way to addressing the reliability issues highlighted in the analysis.

The WCT is therefore recommended for use in the selection process to provide a qualitative assessment of written communication which is superior to the current method that is used. This assessment can be used to identify candidates who need to develop their written communications skills in order to be able to perform train driver written tasks.

The analysis of fairness produced two significant differences in the overall score on version 1 according to age and ethnicity. However, small sample sizes and the amendments that have been made to the revised versions should help towards mitigating some of these issues. In addition, assessment methods can only be said to have adverse impact if there is a potential unfavourable outcome for candidates based on their score. The WCT is not recommended as a pass/fail assessment so cannot have adverse impact if used as recommended.

5.6 SJE

All results presented relate to version A of the SJE. Each of the three main behavioural criteria scores (Conscientiousness, DCS and TLS) and all of the related sub-criteria scores were evaluated.

Another version of the SJE – version B – was also trialled but was discounted at an early stage of analysis due to poor parallel forms reliability.

5.6.1 Sample

A total of 69 participants completed version A of the SJE. Of these, 59 had some job performance data. All of the 59 had manager ratings of behaviour, 43 had driving performance data and 21 had training performance data.

5.6.2 Construct validity

The SJE was designed to provide a valid measure of each of the main behavioural criteria (Conscientiousness, DCS, and TLS) through coverage of the sub-criteria.

5.6.2.1 Convergent validity

Main criteria scores

Pearson correlations were calculated to determine the strength of correlation between the main behavioural criteria as measured by version A of the SJE, and the main behavioural criteria as measured by managers' ratings of behaviour. Please see Table 40 below.

Table 40 - Correlations between SJE version A scores and managers ratings of the three main behavioural criteria

Manager ratings of behaviour		SJE Version A – Main Criteria Scores		
		Conscientiousness	*DCS*	*TLS*
Conscientiousness	*r*	**.33**	.31	.30
	P	.01	.01	.01
	N	59	59	59
DCS	*r*	.33	**.38**	.47
	P	.01	<.001	<.001
	N	59	59	59
TLS	*r*	.17	.23	**.32**
	P	.10	.04	.01
	N	59	59	59

For each of the main criteria, SJE scores and manager ratings of behaviour were significantly correlated as predicted.

Sub-criteria scores

Pearson correlations were calculated to determine the strength of correlation between the sub-criteria as measured by SJE version A, and the sub-criteria as measured by managers' ratings of behaviour. Appendix Section G.9 provides details of all the correlations between SJE version A sub-criteria scores and manager ratings of behaviour.

The analyses showed significant correlations between all SJE version A subscale scores (apart from 'checking') and their related manager ratings. In other words, of the 14 relationships predicted between SJE subscale scores and manager ratings of behaviour, 13 reached significance. Only one correlation would be expected by chance. These correlations ranged from $r(57) = .23$, $p = .042$ to $r(57) = .378$, $p = <.001$.

The correlation between the SJE 'checking' subscale and manager ratings of checking did not reach significance, $r(57) = .02$, $p = .44$. As explained in the development section (Appendix C) steps were taken to address this by enhancing this scale with items from the alternative version of the SJE that was trialled.

In summary, the analyses suggest that SJE version A has very good convergent validity as there was strong evidence that the variables correlated highly with managers' ratings of behaviour.

5.6.2.2 Discriminant validity

Pearson correlations were calculated to determine the strength of correlations between SJE version A main criteria scores. The results are provided in Table 41.

Table 41 - Correlations between SJE version A scores

		SJE Version A – Main Criteria Scores		
		Conscientiousness	*DCS*	*TLS*
Conscientiousness	*r*	**1**		
	P			
	N			
DCS	*r*	.90	**1**	
	P	<.001		
	N	69		
TLS	*r*	.78	.69	**1**
	P	<.001	<.001	
	N	69	69	

Correlations between the three SJE main criteria scores were high (averaging r = .77). These results suggested that the SJE main criteria scales were measuring very similar things.

Pearson correlations were also calculated for the SJE version A scores with all other assessment method scores (please refer to Appendix H for the full results). There was a significant correlation between SJE and MMI conscientiousness scores (which is to be expected as they are different ways of measuring the same construct – for further discussion please refer to the SJE and MMI combined score incremental validity in section 7.4.3.2). The number of significant correlations observed did not exceed the number that would be expected by chance, suggesting that the SJE was measuring something different to the other assessment methods.

Despite the low levels of internal discriminant validity, it is recommended that each scale is retained within the SJE because there is a strong theoretical basis for each scale, the scales have practical use for guiding the MMI, the trials showed that good discriminant validity has been demonstrated with the other assessment methods, and it is not advisable to make such a significant amendments to the SJE based on the relatively small trial sample.

5.6.3 Content validity

The scenarios in the SJE were based on railway operational scenarios that were subsequently translated into an everyday life context. The response options were specifically designed to represent the behavioural attributes that had been defined as part of the train driver selection criteria. Therefore, the evidence suggests that SJE has excellent content validity for the assessment of these selection criteria. For more information on the development of the SJE please refer to Appendix C.

5.6.4 Criterion validity

5.6.4.1 Main criteria scores

Appendix section G.9 provides details of all the correlations between SJE version A main criteria scores and operational performance data. Of the 18 relationships predicted between SJE main scale scores and operational performance, seven reached significance, which is statistically more than would be expected by chance (one significant correlation would be expected by chance).

Notable results included:

- SJE Conscientiousness correlated significantly with:
    - Speeding record ($r(41) = .32$, $p = .02$).
    - Station overruns ($r(41) = .49$, $p = <.001$).
    - Trainee rules assessment ($r(19) = .67$, $p < .001$).
    - Trainee theoretical performance ($r(19) = .63$, $p = <.01$).
    - Trainee practical performance ($r(19) = .59$, $p = <.01$).
    - Manager ratings of overall communications ($r(51) = .36$, $p = <.01$).
- SJE DCS correlated signficantly with manager ratings of overall communication ($r(51) = .30$, $p = .02$).
- SJE TLS correlated signficantly with station overrun record ($r(41) = .56$, $p = <.001$).

5.6.4.2 Sub-criteria scores

For the purposes of informing the MMI, the SJE was also designed to provide a valid measure of each of the behavioural sub-criteria.

Pearson correlations were calculated to determine the strength of correlation between SJE version A sub-criteria scores and operational and training performance. A number of interesting significant correlations were found.

Notable results included:

- SJE ‘commitment to work’ correlated signficantly with the three measures of trainee performance; rules ($r(19) = .69$, $p < .001$), traction ($r(19) = .66$, $p < .001$), and practical ($r(19) = .62$, $p < .001$).
- SJE ‘attention to detail’ correlated signficantly with speeding and overruns performance measures ($r(41) = .27$, $p = .04$ and $r(41) = .39$, $p = .01$ respectively).
- SJE ‘checking’ correlated signficantly with speeding ($r(41) = .35$, $p = .01$).
- SJE ‘rules’ correlated signficantly with speeding ($r(41) = .29$, $p = .03$), overruns ($r(41) = .33$, $p = .01$) and trainee rules ($r(19) = .51$ $p = .01$).
- SJE ‘calm under pressure’ correlated signficantly with handling abnormal events ($r(41) = .31$, $p = .02$).
- SJE ‘sensation seeking’ and ‘need for stimulation’ correlated signficantly with station overruns ($r(41) = .52$, $p < .001$, and $r(41) = .43$, $p < .01$, respectively).

Out of the 210 possible significant correlations, 60 were found. This was significantly more than the 11 correlations that would be expected by chance.

In summary, the analyses suggest that SJE version A has good criterion validity as the evidence showed that the variables correlated with the relevant operational performance ratings as predicted.

5.6.5 Face validity

Qualitative comments collected during the trials from participants and assessors demonstrated clear support for the face validity of the SJE. A key theme arising from the comments was that it was clear to see how the SJE measured constructs that are relevant to the train driver role. A couple of comments were made about specific items being less relevant to the role. This was used with evidence from the trial about the effectiveness of individual items to remove some items from the final version of the SJE.

5.6.6 Reliability

The number of items making up each sub and main criteria scale in version A of the SJE, and the corresponding alpha values, are provided in Table 42.

Table 42 – Items and alpha values for each version A SJE scale

Test measure	Items	Cronbach's Alpha
Main criteria		
Conscientiousness	47	.63
DCS	34	.72
TLS	18	.68
Sub-criteria		
Dependability	20	.69
Attitude to work and people	11	.46
Commitment to work	20	.74
Attention to detail	21	.69
Checking assumptions	13	.70
Rule compliance	19	.70
Calmness under pressure	22	.67
Reactivity to stress	15	.74
Proactive	20	.73
Tenacious	23	.71
Assertiveness	15	.63
Sociability	10	.68
Sensation seeking	6	.70
Need for stimulation	14	.65

The SJE version A main scales were found to be reliable with Cronbach's alphas ranging from 47 items; α = .63 to 34 items; α = .72, all around the α = .7 value specified in the evaluation guidance.

All of the SJE subscales apart from 'attitude to work and people' were found to be reliable with Cronbach's alphas ranging from 14 items; α = .63 to 20 items; α = .74.

For the 'attitude to work and people' scale the original SJE alpha value was low, and so for the final version of the SJE effective items were added from SJE version B. It is anticipated that because these items had good Cronbach's alpha values on version B of the SJE, they will enhance the alphas displayed in Table 42.

In summary, the analyses suggest that SJE version A has good levels of reliability which is likely to be enhanced by the addition of 'attitude to work and people' items from SJE version B.

5.6.7 Fairness

5.6.7.1 Main criteria scores

It was expected that there would be no significant differences in SJE main or sub-criteria scores according to sex, ethnicity or age. Table 43 shows the descriptive statistics for the different protected characteristic groups for each main criterion considered as part of this fairness check.

Table 43 - Descriptive statistics for SJE main criteria scores by protected characteristics groups

			SJE score					
			Conscientiousness		**DCS**		**TLS**	
Protected characteristics		*N*	*Mean Score*	*SD*	*Mean score*	*SD*	*Mean score*	*SD*
Gender	*Males*	60	99.89	9.44	99.79	9.60	100.60	9.87
	Females	9	100.70	13.85	101.42	13.00	96.00	10.54
Ethnicity	*White*	61	100.35	9.89	100.48	9.46	100.09	9.99
	Other	8	97.3	11.12	96.37	13.70	99.26	10.78
Age	*Up to 50 years*	57	100.36	9.97	100.29	10.09	100.05	10.04
	Over 51 years	12	98.31	10.41	98.64	9.85	99.76	10.25
Overall sample		69	100	10	100	10	100	10

There were no significant differences on any of the SJE main or sub- criteria scores according to gender, ethnicity or age.

5.6.8 Administration time and cost

The time required to complete the SJE was, on average, 25 minutes (range from 20 minutes to 40 minutes).

5.6.9 Overall conclusions about the SJE

The analysis suggests that the criterion and convergent validities of the SJE version A are very good. There were a notable number of meaningful significant correlations with performance data at the main and sub-criteria levels, and more than expected by chance.

Significant correlations between SJE main criteria scales and manager ratings of behaviour were found as expected, and each exceeded the r = .2 level specified in the evaluation guidance. All SJE subscales (used to inform the MMI) – barr Checking – correlated significantly with relevant manager ratings.

The intercorrelations between the SJE version A main criteria scores were high, suggesting there may be significant overlap in what is measured by each main scale. This is not an area for concern however, as it is common to find high intercorrelations in tests of this nature.

The results suggest that reliability of the SJE version A is acceptable, and the analysis of fairness produced no significant differences between groups with protected characteristics.

Steps have been taken to improve the reliability of Attitude to work and people (version A Cronbach's alpha = .46), and the validity of the Checking scale by adding the most effective items from version B of the SJE (trialled at the same time), to produce version C (the final version recommended for implementation).

There was no benchmark test for the SJE to use as a point of comparison, but the SJE is recommended for use in conjunction with the MMI to replace the Criterion Based Interview (CBI). For an evaluation of the CBI please refer to Annex 3, Section 6.1.

5.7 MMI

The MMI measures all of the behavioural sub-criteria to produce four scores; Conscientiousness, DCS, TLS and Verbal communication (speaking and listening).

5.7.1 Sample

A total of 92 people sat the MMI, Of the sample, 86 had some form of performance data (all 86 had manager ratings of behaviour, 68 had operational performance data and 29 had training performance data).

5.7.2 Construct validity

5.7.2.1 Convergent validity

Pearson correlations were calculated to determine the strength of correlation between the MMI scores and the manager's ratings of behaviour (Table 44).

Table 44 - Pearson correlations between MMI measurement and managers ratings of the three main behavioural criteria.

Manager ratings of behaviour		MMI Main Criteria Scores		
		Conscientiousness	*DCS*	*TLS*
Conscientiousness	*r*	**.21**		
	P	.03		
	N	86		
DCS	*r*		.02	
	P		.44	
	N		86	
TLS	*r*			-.05
	P			.33
	N			86

MMI and manager ratings of Conscientiousness were significantly correlated in the expected direction $r(84) = .21$, $p = .03$. No significant correlations were found between the MMI and manager ratings of DCS or TLS.

Pearson correlations were calculated to determine the strength of correlation between communication as measured by the MMI (speaking and listening), and the manager's ratings of verbal communication and listening (Table 45).

Table 45 - Pearson correlations between MMI verbal communication score and managers' ratings of communication

Manager ratings of communication		MMI Verbal communication score
Verbal communication	*r*	-.01
	P	.46
	N	85
Listening	*r*	.05
	P	.34
	N	85

The analysis showed that there were no significant relationships between MMI Verbal communication and managers' ratings of verbal communication ($r(84) = -.01$, $p = .46$) or listening ($r(84) = .05$, $p = .34$).

5.7.2.2 Discriminant validity

Pearson correlations were calculated to determine the strength of correlations between MMI main criteria scores. The results are provided in Table 46.

Table 46 - Pearson correlations between MMI main criteria scores

Manager ratings of behaviour		**MMI Main Criteria Scores**		
		Conscientiousness	*DCS*	*TLS*
Conscientiousness	*r*	1		
	P			
	N			
DCS	*r*	.75	1	
	P	<.001		
	N	92		
TLS	*r*	.47	.49	1
	P	<.001	<.001	
	N	92	92	

Correlations between the four MMI main criteria scores were moderate to high (averaging r = .56 excluding Verbal communication, and r = .22 including Verbal communication). These results suggest that the three MMI scales were measuring different but related constructs, and that there will be value in using each of the different scales.

Pearson correlations were also calculated for the MMI scores with all other assessment method scores (please refer to Appendix H for the full results). The number of significant correlations observed did not exceed the number that would be expected by chance, suggesting that the MMI has good discriminant validity and does measure something different to the other assessment methods.

5.7.3 Content validity

The topic areas and behavioural markers used in the MMI were specifically designed to match the behavioural selection criteria for train drivers. Railway subject matter experts were closely involved in the development of the MMI. Therefore, it is anticipated that the MMI has excellent content validity. For more information on the development of the MMI please refer to Appendix C.

5.7.4 Criterion validity

Appendix section G.9 provides details of all the correlations between MMI scores and operational and training performance data. Of the 21 relationships predicted between MMI conscientiousness, DCS and TLS scores and operational/ training performance, five reached significance, which is more than the one correlation that would be expected by chance.

Notable results included:

- MMI Conscientiousness correlated signficantly with operation / isolation of safety systems ($r(66) = -.46$, $p = <.001$), train handling ($r(66) = -.22$, $p = .04$), SPAD record ($r(66) = -.52$, $p = <.001$), and SPAD risk measure ($r(66) = .26$, $p = .02$).
- MMI DCS correlated signficantly with operation / isolation of safety systems ($r(66) = -.29$, $p = .01$).

5.7.5 Face validity

Qualitative comments collected during the trials from participants and assessors demonstrated support for the face validity of the MMI. A key theme arising from the comments was that it was clear to see how the MMI measures constructs that are relevant to the train driver role, and that the approach is more thorough than that used in the CBI.

5.7.6 Reliability

The number of items making up each main criteria scale, and the corresponding alpha values, are provided in Table 47.

Table 47 - Items and alpha values for each MMI scale

MMI Score	Items (behavioural indicators)	Cronbach's Alpha
Conscientiousness	25	.54
DCS	23	.77
TLS	13	.67
Verbal communication	6	.48

The DCS and TLS scales were found to be highly reliable (23 items; α = .77, and 13 items; α = .67 respectively).

The reliability of the Conscientiousness and Verbal communication scales required attention (25 items; α = .54, and 6 items; α = .48 respectively) as the evaluation guidance suggests around α = .7 is preferable. Appendix C describes how the MMI has been amended to address these issues.

As was the case in the trials, interviewers will receive training in the questioning and scoring process (which is based on behavioural indicators), and they will also be subject to periodic reassessments. This should enhance inter-rater reliability,

Following the trials, as part of the refinement process of the MMI, the inter-rater reliability of the MMI was measured using recordings taken of interviews during the trials. On the whole agreement was good, but areas where there was some discrepancy between scores given by different interviewers were noted and this information will be used to improve the training.

5.7.7 Fairness

It was expected that there would be no significant differences in MMI scores according to gender, ethnicity or age. Table 48 shows the descriptive statistics for the different groups for each main criterion considered as part of this fairness check.

Table 48 - Descriptive statistics for MMI main criteria scores by fairness groups

			MMI score							
			Conscientiousness		DCS		TLS		Verbal communication	
Protected characteristics		*N*	*Mean Score*	*SD*	*Mean score*	*SD*	*Mean score*	*SD*	*Mean score*	*SD*
Gender	*Males*	81 (80 for verbal)	3.73	.24	3.75	.18	3.65	.30	3.95	.15
	Females	11	3.83	.10	3.81	.09	3.82	.22	3.89	.30
Ethnicity	*White*	82	3.74	.24	3.75	.17	3.67	.27	3.94	.18
	Other	10	3.80	.13	3.77	.13	3.67	.47	3.97	.11
Age	*Up to 50 years*	81	3.76	.14	3.77	.14	3.68	.30	3.95	.16
	Over 51 years	11 (10 for verbal)	3.61	.56	3.68	.30	3.64	.18	3.88	.27
Overall sample		92	3.74	.23	3.76	.17	3.67	.29	3.94	.17

There was no significant difference on any of the main criteria behavioural scores or the verbal communication score according to gender, ethnicity or age.

5.7.8 Administration time and cost

On average, across the trials, the time required to administer the MMI was approximately 50 minutes, with an additional 10 minutes for scoring. It is expected that this duration will reduce slightly to 45 minutes for administration because the number of topic areas has been reduced from eight to six and follow-up questioning is included by default.

5.7.9 Overall conclusions about the psychometric assessment method

The analyses suggest that construct, content and face validity of the MMI is strong.

The Conscientiousness scale showed good criterion and convergent validity, and the DCS showed good criterion validity. Statistical evidence from this trial for the criterion validity of the TLS and Verbal communication scales is weak.

The reliability of the scales ranged from fair to good.

The results suggested that the DCS and Verbal communication scales need to be revisited to enhance construct validity, and that steps should be taken to enhance the reliability of the Conscientiousness and Verbal communication scales. Appendix C describes how the MMI has been amended to address these issues.

The analysis of fairness produced no significant results and the mean scores were similar in different groups. Therefore, it is not expected that there will be significant differences in the pass rates of different groups if the MMI is implemented.

In comparison to the CBI, the initial data suggested that the SJE, in particular, has better validity and reliability than the CBI (information on the reliability of the CBI is limited).

The combination of the SJE and MMI also better addresses good practice as it provides two measures for each behavioural criterion, includes a measurement of assertiveness (as a sub-scale) and is less likely to discriminate against high-potential candidates with limited previous experience (due to the inclusion of situational questions).

6 Further evaluation of current psychometric assessment methods

6.1 Introduction

The purpose of undertaking further evaluation of the current psychometric assessment methods was twofold: (1) to verify that the results of T340 (RSSB, 2005) were still valid, and (2) to compare the evidence of the proposed methods with that of the current methods.

The T340 project identified a number of weaknesses in the existing assessment centre methods. This evaluation was based on a review of the assessment centre process and such previous validation studies as could be accessed. The weaknesses included evidence of low or inconsistent validity coefficients for some of the assessment methods and practical problems with administering and scoring some of the tests, particularly related to those tests that are computer delivered.

Specifically, T340 determined that the evidence for the validity of TRP was acceptable and that the method should be retained. The results were generally considered unacceptable for the DTG and it was recommended that this method should be replaced. The results were inconsistent for the Group Bourdon and CBI and it was recommended that these methods should be either replaced or upgraded.

Since the T340 study, a number of other validation studies have been conducted, mostly by OPC who have provided a summary of these studies to RSSB. These studies were reviewed in order to consider whether this new evidence presents information which might change the T948 project team's view of the technical quality of

the existing assessment methods and how the new methods compare to the existing methods. In addition, RSSB conducted a 'health check' of the existing psychometric assessment process and these results are also taken into account.

All the evidence available on the current assessment methods from the above sources was compared to the evidence relating to the proposed assessment methods which were trialled as part of T628 (RSSB, 2010) and T948. This extensive analysis was undertaken to ensure that informed recommendations could be made about the most suitable methods for use in the future psychometric assessment process.

6.2 Method

A short description of the various sources of information that were used for the review of the current methods is provided below with an outline of the method used for processing/analysing each one.

6.2.1 T340 study

In 2005, a number of companies were approached who had expressed an interest to be closely involved in the validation study. The aim of the data collection methodology was to collect data on as many drivers as possible who had been recruited since 1999 and particularly those recruited since 2003. Performance data were collected for 373 drivers which covered train handling, safety records, following procedures, work attitude and training records. Detailed training records were also collected for 157 drivers. The performance data were then matched to assessment centre records to analyse the strength of the relationship between the assessment methods and operational performance and training data.

6.2.2 RSSB 'health check'

In 2011, RSSB undertook a 'health check' of the train driver psychometric assessment process on behalf of the Driver Selection Governance Group. The purpose of the health check was to determine how well the current process is working. Information that is relevant to this report includes the pass rates and fairness of the various methods.

The health check involved the analysis of candidate data from the RACF industry database (provided by Springfield Training) between 1st April 2010 and 31st March 2011. The analysis was based on 1,968 train driver assessments recorded in the RACF database during the period of interest. Statistical analysis was based on all of the available assessments where full or partial assessments with a final result were recorded.

However, because not all candidates sit all tests the number of valid results varies depending on the analysis conducted. The number of assessments included in the analysis is presented for each assessment method in the following sections.

6.2.3 Evidence provided by OPC

Since the T340 project in 2005, a number of validation studies have been conducted on the current assessment methods, mostly by the Occupational Psychology Centre (OPC) who have provided a summary of these studies to RSSB. A basic meta-analysis was undertaken on these validation studies to determine if there was any evidence which might change the T948 project team's view of the technical quality of the current assessment methods and how the new methods compare to the existing methods.

There were several methodological issues related to the task of reviewing the validation evidence and further details are available on request from RSSB. However, some of the main challenges were that it was difficult to tell from the summary report exactly how many studies were reported. As far as possible, it has been determined which tests, test scores and criteria have been used in each study and all the statistics reported were expressed as correlations so that they could be combined and averaged. However, in some cases the number of tests and criteria may have been either under- or over- estimated. The averaging process was restricted to those significant test–criterion relationships which have been reported. That is, that if a result has been reported in the OPC summary, a claim is being made implicitly that the relationship is meaningful and it should appear in all the studies where that criterion measure was used. In addition to calculating an average correlation for each test score, an estimate of the confidence interval was calculated for each average correlation.

Where the confidence interval is wide, it is an indication that there is a wide range of values for the validities and, therefore, the results are inconsistent. For this data set, confidence intervals smaller than ± 0.05 indicate relatively consistent results while confidence intervals greater than ± 0.10 indicate inconsistent results.

If the confidence interval includes the value of zero for the average correlation, it means that the average correlation is not significantly different from zero. So, for example, if the reported average correlation were 0.6 and the confidence interval were ± 0.08, the confidence interval runs from -0.02 to +0.14 which includes zero and indicates the average value is not significantly different from zero.

6.2.4 T948 study

Finally, as part of the T948 research, assessment centre data were extracted from the RACF industry database for the participants who took part in the trials. This provided test score data for the existing assessment methods that could be correlated with job performance data to assess criterion validity using the same method as for the new assessment methods that were trialled. Statistical analysis was based on all of the available assessments where a final result was recorded. This included data on candidates who passed the assessment and are existing drivers or driver trainees and also

candidates who failed the assessment centre but who agreed to take part in the trials. The number of assessments included in the analysis is presented for each assessment method in the following sections.

6.3 Findings

It should be noted that recommendations were formulated by weighing up all available evidence from the above sources of information. Where there are contradictory results between studies, more weight was given to studies which were considered more robust based on the recency of the information, the quality of the research and the sample sizes.

6.3.1 Group Bourdon

6.3.1.1 Overview

The Group Bourdon is currently used by rail assessment centres to assess attention. It involves candidates identifying and marking examples of a particular pattern of dots which are embedded in a set of other dot patterns. There is a paper and pencil version and a computer-based version of the method but the two versions produce different scores. The paper version can produce a large number of scores but only three scores are actually used by the Rail Assessment Centres which are 'total productions', 'total omissions' and 'total faults'.

The results in the following sections relate to the paper version of the method unless stated otherwise as this version has better validity and is more widely used by assessment centres.

6.3.1.2 Sample

T340 review

Out of the total sample of 373 drivers, performance data were collected for 213 drivers and trainees who completed the paper version of the Group Bourdon and 43 people who completed the computerised version.

RSSB Health check

The health check analysed candidate data from the RACF industry database for 1562 candidates who completed the paper version of the Group Bourdon between 1st April 2010 and 31st March 2011.

Review of OPC summary evidence (2012)

Seven different studies are reported where the Group Bourdon has been evaluated as shown in Table 49. Most of the studies are small with data available for fewer than 100 participants, although two studies (1992 and 1996) have large samples.

Table 49 – Studies included in the OPC summary of evidence relating to the paper Group Bourdon

Year of research	Sample	Type of study
1992	280	A predictive validation study was undertaken with British train drivers. Information was also obtained on their practical train handling performance during training.
1996a	1600	A large scale predictive validation study was undertaken with British train drivers and data was collected on their safety performance.
1996b	200+	A national predictive validation study was undertaken with train drivers. Information was obtained on their competence from driver managers.
2007a	60	A predictive validation study was completed for train drivers and performance data was collected. This study only collected data on the computerised version of the test.
2007b	30	A predictive validation study was undertaken for a British train operating company. Job performance measures were collected from driver mangers including information on safety performance.
2009	140	A predictive validation study was conducted on behalf of a train operating company. Measures of training and job performance were obtained.
2011	190	A predictive validity study was undertaken on 100 trainees and 90 train drivers. Information was later collected on training and job performance.

T948 trials

The Group Bourdon paper version scores were downloaded from the RACF industry database for 137 of the participants who took part in the trials. Detailed information on the demographic characteristics of the sample is provided in Appendix D. Out of these participants, operational performance data were available for 38 and training data were available for 17 participants which were used to assess the criterion validity of the method.

6.3.1.3 Criterion validity

T340

The results of the T340 study found that the relationship between the Group Bourdon and operational performance data had moderate predictive validity, although there was some inconsistency with some of the scores.

The findings for the paper and pencil version of the Group Bourdon were mixed and differed in important ways from the computerised version. The pass rates were very different and the validation evidence showed different results for the computerised version and paper version of the method. Only the paper version was regarded to have validity and utility and this is described in the sections below.

The evidence of criterion validity for the scores in the paper version of the method was mostly acceptable. The three scores used by assessment centres are 'total productions', 'total omissions' and 'total faults' and the criterion validity of these is reported.

- The 'total omissions' score correlated significantly with procedure based performance ($r = .33$, $p < .05$) and classroom training assessment ($r = .40$, $p < .01$).
- The 'total faults' score correlated significantly with safe performance ($r = .31$, $p < .05$) and classroom training assessment ($r = .38$, $p < .05$).
- The 'total production' score did not correlate significantly with any aspect of performance.

These results suggest that the omissions and faults scores are worth considering but the evidence is not present for the productions score.

Review of OPC summary evidence

Table 50 summarises the average criterion validity on the paper version of the Group Bourdon from the eight validation studies that were undertaken between 1992 and 2011. Correlations that met acceptable levels are marked in bold, **green** text.

Table 50 - Criterion validity summary for the paper Group Bourdon

Paper Group Bourdon score	Training: Average Validity	Performance: Average Validity	Largest observed correlation
Part 4 production	-0.03 to 0.03	-0.03 to 0.03	0.00
Part 5 production	-0.031 to 0.03	-0.01 to 0.13	0.12 (performance)
Total production	0.03 to **0.25**	0.04 to 0.15	0.29 (training)
Part 4 omissions	-0.03 to 0.03	0.03 to 0.11	0.11 (performance)
Part 5 omissions	-0.03 to 0.03	-0.03 to 0.11	0.10 (performance)
Total omissions	0.01 to 0.14	0.03 to **0.26**	0.41 (performance)
Group omissions	-0.03 to 0.03	-0.03 to 0.11	0.31 (performance)
Total faults	0.02 to 0.14	-0.04 to 0.07	0.22 (performance)

As these averaged correlations are based on eight studies, it could be argued that more weight should be given to these results as the combined sample sizes are much larger than that of the T340 study. The results clearly indicate that there is acceptable evidence of criterion validity for the 'total productions' score which correlates with training outcomes, and for the 'total omissions' score which correlates with operational performance. The confidence interval of the 'total faults' score runs from -0.04 to 0.07 which includes zero and indicates the average value is not significantly different from zero. The results for 'total faults' did not replicate the results found in the T340 study which suggests that this particular score may be unstable.

The other scores were considered even though they are not used in the assessment decision, out of interest. It was not clear from the summary of the studies whether the Part 4 and Part 5 scores were always evaluated and so their validities may be artificially deflated but, in any case, the correlation between these scores and total scores is so high ($r > 0.92$) that there is little value in recording them. Furthermore, the correlation between the 'total omissions' and 'group omissions' score is also very high ($r = 0.88$), so it would seem that these are not worth considering separately. The average validities suggest that the predictive power of these scores is low but non-zero. This coincides with the findings of the T340 study (RSSB, 2005) which suggested that there were some good correlations between some Group Bourdon scores but not all of them. It appears that there may only be utility in retaining the 'total productions' and 'total omissions' scores.

T948 trial results

The Group Bourdon scores collected as part of the T948 trial were correlated with operational performance and training assessments. Table 152 in Appendix G provides a record of these results. Before the analysis was undertaken, predictions were made about the expected relationships between the scores and performance measures as a sense check (see appendix F). Out of the 36 relationships predicted, only one correlation was significant and in the expected direction (four were significant but in the wrong direction).

This correlation was between 'total productions' and speeding ($r(38) = .32$, $p = .03$). This coincides with the evidence found in the previous validation studies provided by OPC although the score correlated with operational performance in this study, rather than training assessments.

6.3.1.4 Content validity

There is not a test manual available for the Group Bourdon so it was difficult to assess the content validity of the method. However, at a superficial level, the method appears to be an adequate measure of selective visual attention as candidates are required to respond to certain stimuli (patterns of dots) whilst ignoring others. The method is very similar to the Differential Attention test (DAKT) from the Vienna Test System which also claims to be a measure of selective attention.

6.3.1.5 Construct validity

Convergent validity

It is possible that the Time measure of the computerised Group Bourdon is a measure of information processing speed and that being able to take in and process information quickly would help during training and learning of procedures. This hypothesis is supported by the finding that the Time score correlates in the range 0.3 to 0.39 with the various versions and parts of the TRP and above 0.50 with the DTG Good scores.

Discriminant validity

In order to assess the discriminant validity of the Group Bourdon, the three main scores were correlated with scores from other assessment methods as part of the T948 trials. 'Total productions' was significantly correlated with DTG part 3 good ($r(124) = .35$, $p < .001$) and DTG self-paced wrong ($r(77) = .31$, $p = .01$). These correlations are low to moderate and may be due to the speed at which candidates complete the tests.

The omissions and faults scores did not significantly correlate with any other assessment methods in the current test battery or the

trialled tests. This suggests that the Group Bourdon scores have added value over and above the other assessment methods.

In consideration of the previous validation studies provided by OPC, there is quite good evidence for the differential validity of the three main scores. Pearson correlations were calculated to determine the size of correlations *between* Group Bourdon test scores and the results are presented in Table 51.

Table 51 – Pearson correlations between Group Bourdon scores

		Production total	Omissions total	Faults total
Production total	*r*		-.004	.003
	P		.965	.977
	N		137	137
Omissions total	*r*	-.004		-.049
	P	.965		.572
	N	137		137
Faults total	*r*	.003	-.049	
	P	.977	.572	
	N	137	137	

There were no significant correlations between the three main scores which suggests that there is utility in considering each of them, provided that the evidence of criterion validity is sufficient.

6.3.1.6 Face validity

The importance of the underlying constructs of concentration and perceptual speed is recognised as appropriate but the precise nature of the task and its relevance to real world tasks is less obvious. It is not known from RSSB research whether candidates consider the Group Bourdon to have face validity, although OPC has remarked in a report to the HSE (Fletcher, 2004), that the similar SCAAT test does not have face validity for train driver candidates.

6.3.1.7 Reliability

It was difficult to assess the reliability of the Group Bourdon as there was insufficient evidence due to there not being a test manual for this assessment method. However, whilst there is no direct evidence for the reliability of the test, it was possible to estimate the reliabilities from the correlations between the different parts of the test using the previous validation trials provided by OPC.

This analysis indicated that the 'total production' score has very good reliability, ranging from 0.82 to 0.96. However, the reliabilities for the omissions and faults scores may be less good, and in some cases were estimated to be as low as 0.21. As noted previously, as the Group Bourdon is very similar to the DAKT test, the reliabilities reported for the production score is considered a good estimate.

A related concern is the reliability of scoring of the paper and pencil version of the Group Bourdon. No evidence was presented on this point in any of the previous validation reports but the error rate in hand-scoring similar tests is known to sometimes be quite high.

6.3.1.8 Fairness

In order to assess the fairness of the Group Bourdon, results from the RSSB health check (RSSB, 2011), the previous validation studies and the current validation study were reviewed.

RSSB health check

The RSSB health check reviewed the completed driver assessments on Group Bourdon scores to determine if there were significant differences between candidates according to gender, ethnicity and age. The results are shown in Table 53.

Table 52 - Descriptive statistics for Group Bourdon scores by fairness groups

			Group Bourdon scores					
			Productions total		**Omissions total**		**Faults total**	
Protected characteristics		*N*	*Mean score*	*SD*	*Mean score*	*SD*	*Mean score*	*SD*
Gender	*Males*	1564	116.38	385.81	20.36	21.76	.67	4.71
	Females	117	1218.50	239.94	18.46	23.82	.57	1.34
Ethnicity	*White*	1305	1157.53	410.53	17.93	18.01	.58	5.06
	Other ethnicity	371	1208.57	226.77	28.42	30.80	.59	1.98
Age	*Up to 50 years*	1584	1175.03	385.40	20.38	22.25	.66	4.68
	Over 51 years	86	1061.95	187.12	16.79	13.77	.66	1.58
Overall sample		1681	1169.08	377.68	20.22	21.91	.66	4.56

There were no significant differences on any of the Group Bourdon scores according to gender. However, there were significant differences on the total production and total omissions scores according to ethnicity:

- There was a significant difference on total productions ($t(1674) = -2.29$, $p = <.05$, where white participants scored slightly better than non-whites.
- There was a significant difference on total omissions ($t(444) = -6.27$, $p = <.001$, where white participants scored slightly better than non-whites.

There were also significant differences on the total production and total omissions scores according to age:

- There was a significant difference on total productions ($t(1668) = 2.70$, $p = <.01$, where participants aged 21 to 50 scored significantly better than older candidates aged 51+.
- There was a significant difference on total omissions ($t(110) = 2.27$, $p = <.03$, where participants aged 21 to 50 scored significantly better than older candidates aged 51+.

The RSSB health check also reported issues with the fairness of the Group Bourdon with the current pass criteria. The four-fifths rule was failed for both black and Asian candidates on the paper version. It was not possible to assess the computerised version of the method as the sample was too small and the pass rate was 99% so any analysis would have been meaningless. Table 53 illustrates the difference in pass rates between ethnic groups. The pass rates were significantly lower for black candidates (Chi-square = 26.666, df = 1, p < 0.001) and Asian candidates (Chi-square = 7.237, df = 1, p = 0.007) than for white British candidates. These results demonstrated that the pass rates for black and Asian candidates were marginally less than 4/5ths of the pass rate for white British candidates.

Table 53 - Pass rates for the paper Group Bourdon according to ethnic group Ethnic group

	Pass (%)	Fail (%)	No. Assessments in analysis
Black or black British	41	59	192
Asian or Asian British	48	52	125
White - British	60	40	1245
Totals	57	43	1562

OPC summary review

According to the studies provided, there was some evidence of differences in response related to age, gender and ethnicity for the Group Bourdon though these effects were generally small and did not consistently favour one group over another. No details were

provided regarding the way that this conclusion was reached so it was not possible to judge why these results are so inconsistent with the findings of the RSSB health check.

6.3.1.9 Administration time and cost

The test is easy to administer but the scoring of the paper and pencil version is a little time consuming. There may also be problems with scoring errors but the extent of this is unknown. The current hardware and software used for delivering the computer-based version is old, unreliable and poorly supported which is why the computerised version is not being considered for future use.

6.3.1.10 Overall conclusion about the method

There is substantial evidence that the computerised version and paper version of the Group Bourdon are not equivalent in scoring or pass rates. Data for the computerised version were very limited so only the paper version has been considered in depth. The computerised version is not recommended for continued use as part of the train driver assessment process.

There have been over ten studies of the Group Bourdon which provides a rich source of evidence to consider the merits and limitations of the method. These studies include eight validation studies from OPC, two validation studies from RSSB and the recent health check. More weight was given to the review of validation studies from OPC due to the combined large sample sizes and the fact that the validity coefficients were averaged with confidence intervals. These results indicated that there are two scores worth considering in the Group Bourdon and these are 'total production' and 'total omissions'. These scores showed acceptable validity coefficients with training assessment and operational performance. The confidence interval of the 'total faults' score indicated the average value was not significantly different from zero. Therefore, there seems to be merit in using only the 'total production' and 'total omissions' scores to assess selective attention.

There is evidence, mainly from the RSSB health check, that the current scoring rules result in significantly different pass rates for different ethnic groups. The four-fifths rule was failed for black and Asian candidates. It is however, possible to reduce the pass marks on the Group Bourdon so as to reduce between group differences and improve the fairness of the method to acceptable levels. This is discussed further in section 7 in Annex 3.

A further consideration is that attention has been defined with two sub-criteria: selective and divided attention. The Group Bourdon provides a measure of selective attention but does not assess divided attention.

It is therefore recommended that the paper version of the Group Bourdon could be considered in conjunction with another measure of attention provided that the pass criteria are changed to reduce between group differences in ethnic groups.

6.3.2 DTG

6.3.2.1 Overview

The DTG is currently used by rail assessment centres and is intended to give measures of the ability to react safely and quickly and eye-hand-foot coordination. It is a computer administered test which requires candidates to respond to on-screen shapes in different colours and to different sounds by pressing buttons or pedals. The test has three phases, a practice phase, a timed performance phase and a self-paced phase.

6.3.2.2 Sample

T340

This study collected performance data for drivers who had completed the DTG in previous assessments. Performance data was collected for 255 drivers who had completed the DTG.

RSSB Health check

The health check analysed candidate data from the RACF industry database for 775 candidates who completed the DTG between 1st April 2010 and 31st March 2011.

Review of OPC summary evidence

Nine different studies were reported where the DTG has been evaluated as shown in Table 54. Three of these studies had moderate to large sample sizes. The remainder had small samples.

Table 54 - Studies included in the OPC summary of evidence relating to the DTG

Year of research	Sample	Type of study
1992a	280	A predictive validation study was undertaken whereby performance data on practical handling were obtained following the assessment centres had selected candidates.
1992b	200	A predictive validation study was undertaken with train drivers and SPAD records were obtained.
1996a	200	A predictive validation study was undertaken and competence information was collected from driver managers.
1996b	1600	A large scale validation study was undertaken within Great Britain and operational safety performance data was collected on existing train drivers who had passed the DTG previously in an assessment centre.
2002	50	A concurrent validation study was conducted on existing drivers and data was collected on safety on the job.
2007a	90	A predictive validation study was conducted for an overseas rail operator that used the DTG. Performance data was collected from driver managers on safety performance.
2007b	30	A predictive validation study was undertaken for a British rail operator. Again, performance data was collected on safety records.
2009	130	A predictive validation study was undertaken within a British rail operator. Training assessments and performance records were collected.
2011	100	A predictive validation study was undertaken within a British rail operator. Training assessments and performance records were collected.

T948 trials

A total of 164 participants completed the DTG as part of their original driver assessment. Detailed information on the demographic characteristics of the sample is provided in Appendix D. Out of the 164 participants who completed the DTG, 64 had operational driving performance data and 25 had training performance data.

6.3.2.3 Criterion validity

T340

The results of the T340 study found that the relationship between the DTG and operational performance data was weak and there was some inconsistency with some of the scores.

The three scores used by assessment centres are 'part 3 good', 'part 3 omissions' and 'self-paced wrong' and the criterion validity of these scores is reported.

- Part 3 good was significantly correlated with procedure based work ($r = .16$, $p < .05$).
- Part 3 omissions was significantly correlated with procedure based work ($r = .14$, $p < .05$).
- Self-paced wrong was significantly correlated with safe performance ($r = .13$, $p < .05$).

These validity coefficients fall below 0.2 which is considered by EFPA to be the minimum criterion. None of the reported correlations for the three scores which are used met this acceptable criteria. Further, given that the DTG is intended to provide a measure of the ability to react safely and quickly and provide a measure of eye-and-foot coordination, the relationship with procedure based work is not one that would relate directly to these skills. Further, there was no evidence of a significant relationship with training assessments. Both this and previous validation studies suggest that the DTG's predictive validity for the key areas of train handling is low though not zero.

Review of OPC summary evidence

Table 55 summarises the average criterion validity on the DTG from the nine validation studies that were undertaken between 1992 and 2011.

Table 55 - Criterion validity summary for the DTG

DTG score	Training: Average Validity	Performance: Average Validity	Largest observed correlation
Part 3 good	0.01 to 0.22	0.01 to 0.17	0.31 (performance)
Part 3 omissions	-0.04 to 0.04	-0.01 to 0.01	0.00
Self-paced wrong	-0.04 to 0.04	-0.01 to 0.01	0.00

These results indicate that the omissions and self-paced wrong scores have no predictive value. There were some moderate validity coefficients for the 'part 3 good' score but the average validity was low. These findings are in line with the T340 validation study (RSSB, 2005) where there was little evidence in support of the DTG.

T948 trial results

Table 153 in appendix G provides details of all the correlations between DTG test scores performance data. Of the 27 relationships predicted between DTG test scores and operational performance (see Appendix F), there were no significant correlations. This aligns with the previous validation studies and demonstrates the DTG scores have little predictive value.

6.3.2.4 Content validity

There is not a test manual available for the DTG so it was difficult to assess the content validity of the method. However, at a superficial level, whilst the actions of responding to visual stimuli are related to the tasks that train drivers undertake, the eye-hand-foot coordination required in the test is not representative of the driver tasks.

6.3.2.5 Construct validity

Convergent validity

DTG scores were significantly correlated with Group Bourdon and TEA-Occ scores in the region of r = -.37 to .40. This suggests that the DTG requires the candidate to make use of attentional capacity. The DTG scores were also significantly correlated with TRP part two (r = .21 to .23) which may be due to the likely relation with general intelligence.

Discriminant validity

The review of OPC validation studies showed that the correlations between the various DTG scores tend to be high (0.61 to 0.81) suggesting there is little added value in the different scores, except for the self-paced wrong scores which have low correlations with the other measures but little evidence of validity.

As an additional check, the DTG scores from the T948 trials were correlated. The results are provided in Table 56 below.

Table 56 – Pearson correlations between DTG scores

		Part 3 good	Part 3 omissions	Self-paced wrong
Part 3 good	*r*		-.53	-.17
	P		<.01	.09
	N		163	95
Part 3 omissions	*r*	-.53		.06
	P	<.01		.55
	N	163		94
Self-paced wrong	*r*	-.17	.06	
	P	.09	.55	
	N	95	94	

These results demonstrated that the 'part 3 good' and 'part 3 omissions' score were significantly correlated ($r(163) = -.53$, $p = <.01$). The 'self-paced wrong' score did not significantly correlate with the other two scores. Given the combined results of the available validation studies, it appears that there is little added value in the different scores, except for the 'self-paced wrong' but this score appears to have no predictive validity.

6.3.2.6 Face validity

The test has some face validity and a case can be made for its relationship to tasks and actions undertaken by train drivers. On the other hand, it is not clear that the speed of reaction and the nature of the eye-hand-foot coordination required in the test are that similar to current role requirements.

6.3.2.7 Reliability

Evidence for the reliability of the DTG scores suggests that they are in the range good to very good. Estimates produced in the CAS report suggest reliabilities in the range 0.78 to 0.90. Information for other very similar tests, for example the Determinations Test (DT) from the Vienna Test System suggests that the reliabilities may be even higher, in the range 0.92 – 0.96.

6.3.2.8 Fairness

In order to assess the fairness of the DTG, results from the RSSB health check (RSSB, 2011), the previous validation studies and the current validation study were reviewed.

RSSB health check

The RSSB health check reviewed the completed driver assessments on DTG scores to determine if there were significant differences between candidates according to gender, ethnicity and age. The results are shown in Table 57.

Table 57 - Descriptive statistics for DTG scores by fairness groups

			DTG scores					
			Part 3 good		Part 3 omissions		Self-paced wrong	
Protected characteristics		*N*	*Mean score*	*SD*	*Mean score*	*SD*	*Mean score*	*SD*
Gender	*Males*	761	70.22	24.07	9.07	10.35	12.24	14.50
	Females	61	66.95	27.05	8.48	9.89	12.39	14.69
Ethnicity	*White*	703	71.13	23.99	8.76	10.16	11.36	13.68
	Other ethnicity	115	63.26	24.76	10.46	10.99	17.70	18.07
Age	*Up to 50 years*	774	71.19	23.73	8.84	10.27	12.10	14.52
	Over 51 years	43	48.72	24.91	12.42	10.60	14.49	14.58
Overall sample		822	69.98	24.30	9.02	10.31	12.25	14.50

There were no significant differences on the DTG scores according to gender. However, there were significant differences according to ethnicity and age. The differences found on ethnicity are as follows:

- There was a significant difference on part 3 good ($t(816) = 3.24$, p = <.001, where white participants scored slightly better than non-whites.
- There was a significant difference on self-paced wrong ($t(136) = -3.99$, p = <.001, where white participants scored significantly better than non-whites.

The differences found on age are as follows:

- There was a significant difference on part 3 good ($t(815) = 6.03$, p = <.001, where participants aged 21 to 50 scored significantly better than older candidates aged 51+.

- There was a significant difference on self-paced wrong (*t*(815) = -2.22, p = <.03, where participants aged 21 to 50 scored significantly better than older candidates aged 51+.

The RSSB health check expressed some concerns with the fairness of the DTG with the current pass criteria. The four-fifths rule was not failed for ethnic groups but the pass rates were significantly lower for black candidates (Chi-square = 9.457, df = 1, p = 0.002) and Asian candidates (Chi-square = 6.761, df = 1, p = 0.009) when compared to the pass rates for white British candidates, as shown in Table 58.

Table 58 - Pass rates for the DTG according to ethnic group Ethnic group

	Pass (%)	Fail (%)	No. assessments in analysis
Black or black British	70	30	50
Asian or Asian British	71	29	42
White - British	86	14	683
Totals	84	16	775

Review of OPC summary evidence

The review of the nine validation studies found significant age differences, with younger candidates scoring better than older ones.

This review also found that ethnic minorities score lower than white candidates on average on the good scores but the differences are quite small and do not fail the four-fifths rule. No gender differences were found. Concern has been expressed in the past that experienced railway staff (eg guards) would be disadvantaged by the test but there is no evidence to support that concern.

6.3.2.9 Administration time and cost

The DTG is computer administered and it takes 15 minutes to complete. It requires a computer and screen, a special keyboard and pedal set. The test is easy to administer being delivered and scored on computer. However, the current hardware and software is old, unreliable and poorly supported according to feedback from rail assessment centres.

6.3.2.10 Overall conclusion about the method

The purpose of undertaking further evaluation of the current psychometric assessment methods was to verify that the results of T340 were still valid. The results of the nine additional validation studies provided by OPC, the T948 study and the RSSB health check were consistent. That is, there was little evidence of the predictive validity of the DTG with operational performance or

training assessments. There was some evidence of weak criterion validity for the 'self-paced wrong' score but this is low though not zero.

In addition to the validation findings, another issue with the DTG which was reported by rail assessment centres was that the current hardware and software is old, unreliable and poorly supported. With regard to any concerns with replacing the DTG, it should be noted that it seems possible to predict scores on the DTG with a high degree of accuracy using a combination of scores from the new tests. This means that the information being provided by the DTG is replicated by a combination of other methods.

In summary, it is recommended that the DTG should be replaced as there is sufficient evidence to indicate that the method has very limited validity and there are severe problems with the lack of technical support that is provided to assessors.

6.3.3 Trainability for Rules and Procedures test (TRP)

6.3.3.1 Overview

The Trainability for Rules and Procedures test (TRP) consists of two parts, which are known as TRP1 and TRP2.

In TRP1, candidates listen to a recording of fictitious safety-related rules information. They learn the information and then answer 18 questions from memory based on the information they have just heard and read. It is intended to provide a measure of verbal memory and comprehension.

In TRP2, candidates are presented with fictitious dials for a train cab and rules relating to their functioning. Candidates are required to apply the rules to decide the order in which the dials should be checked. There are 43 sets of dials for candidates to work through. TRP2 is intended to provide a measure of rule application and non-verbal reasoning.

6.3.3.2 Sample

T340

The TRP assessment method was updated before this validation study took place. Therefore, scores on the recent version that is used today were only available for 43 candidates in the sample.

RSSB Health check

The health check analysed candidate data from the RACF industry database for 1398 candidates who completed TRP1 and 1401 candidates who completed TRP2 between 1st April 2010 and 31st March 2011.

Review of OPC summary evidence

Four different studies are reported where the TRP has been evaluated as shown in Table 59. Sample sizes were relatively small, with the exception of the 2011 study.

Table 59 – Studies included in the OPC summary of evidence relating to the TRP 1 and TRP 2

Year of research	Sample	Type of study
1993	35	A national concurrent validity study was undertaken. Trainees in training were invited to complete the TRP and at the same time training measures were collected for Trainman D Schedule 1 training.
2001	50	A national (British) concurrent validity study involving a number of train operators. About 50 existing train drivers were invited to complete a range of tests and exercises. At the same time job and safety performance measures were collected.
2009	30-40	A predictive validation study undertaken on train drivers within a British train operator.
2011	134	Validation study undertaken on train drivers within a British rail organisation.

T948 trials

A total of 175 participants completed the TRP as part of their original driver assessment. Detailed information on the demographic characteristics of the sample is provided in Appendix D. Out of these participants, 65 had operational driving performance data and 25 had training performance data.

6.3.3.3 Criterion validity

T340

This review concluded that there is strong support for the predictive validity of the Trainability for Rules and Procedures test (both parts one and two). There were significant relationships to training outcomes, safe performance and procedure based work as shown in Table 60. All the significant correlations found were predicted relationships.

- TRP part one was significantly correlated with procedure based work ($r = .30$, $p < .01$), safe performance ($r = .35$, $p < .05$) and classroom training assessment ($r = .23$, $p < 0.5$),
- TRP part two was significantly correlated with safe performance ($r = .22$, $p < .01$) and classroom training assessment ($r = .18$, $p < .05$).

These results suggest that TRP1 may be a better predictor than TRP2 but this may be due to the small sample sizes for this study (N = 43). It is anticipated that TRP2 would have at least as good predictive relationships with operational performance as the old

version of TRP2 since the main change is in how the test is scored rather than in the content.

Review of OPC summary evidence

Table 60 summarises the average criterion validity on the TRP1 and TRP2 from the four validation studies that were undertaken between 1993 and 2011.

Table 60 - Criterion validity summary for the TRP

	Training: Average Validity	Performance: Average Validity	Largest observed correlation
TRP part 1	**0.26** to **0.41**	-0.02 to 0.02	0.41 (training)
TRP part 2	0.15 to **0.26**	0.15 to **0.25**	0.41 (training)

The TRP appears to be one of the more valid tests in the current assessment centre process. The T340 study findings that the TRP1 was a good predictor of operational performance was not replicated in these validation studies. However, the TRP1 score did have acceptable levels of criterion validity with training assessments.

The results demonstrate that TRP2 has predictive validity with both training assessments and operational performance which supports the results of the T340 study. This is consistent with information provided by OPC on the separate validation studies which confirms that the TRP correlates with a range of training measures, but the evidence of criterion validity with job performance measures is less certain.

T948 trial results

The criterion validity of TRP1 and TRP2 are reported separately. Table 154 and Table 155 in Appendix G provide details of all the correlations between TRP parts 1 and 2 with operational and training performance data.

TRP1

Of the six relationships predicted between TRP part one and operational performance, there were no significant correlations in the expected direction (one was significant in the wrong direction).

TRP2

Of the 14 relationships predicted between TRP2 and operational performance, there were no significant correlations in the expected direction (one was significant in the wrong direction).

Despite these poor findings, the validation evidence from the T948 trials should not be given much weight in the overall conclusion about the TRP as the sample size was very small and ranged from 25 to 65 participants which is a methodological limitation.

6.3.3.4 Content validity

The TRP assessment method is relevant to some aspects of the train driving task in terms of its relation to instrument interpretation.

6.3.3.5 Construct validity

OPC have not produced a test manual for the TRP as the use of this test is described within the Train Driver's Procedures Manual in terms of administering, scoring and interpreting results. As such, it was difficult to assess the theoretical construct of this test fully although it does seem to have clear links with the ability to learn training related information.

Discriminant validity

The review of previous validation studies provided by OPC found that the two parts are relatively uncorrelated (0.21) so there should be added value in using both test scores.

As part of the T948 trials, the TRP scores were correlated with other tests. TRP1 was significantly correlated with TAVTMB overview ($r(85) = .23$, $p = .03$) but it was not significantly correlated with any other test scores in the current test battery or the trialled tests which suggests that it would provide unique information in the assessment process.

TRP2 was significantly correlated with VIGIL ($r(55) = .29$, $p = .03$), TEA-Occ dual task decrement ($r(92) = -.34$, $p < .001$), DTG part 3 good ($r(162) = .21$, $p = .01$) and DTG self-paced wrong ($r(93) = -.23$, $p = -.03$). Whilst there are more correlations, these were all low to moderate which suggests that other tests only provide a partial measure of non-verbal reasoning.

6.3.3.6 Face validity

The face validity of this test is good and is relevant to the train driving role. Candidates have reported that they believe the TRP to have good face validity with instrument interpretation.

6.3.3.7 Reliability

Reliability was assessed by referring to the previous validation studies provided by OPC.

The reliability of the TRP1 was assessed using Cronbach's alpha ($\alpha = .662$) and split half reliability ($\alpha = .614$).

The reliability of TRP2 is ($\alpha = .955$) and split half reliability is ($\alpha = .877$).

These are acceptable levels of reliability.

6.3.3.8 Fairness

RSSB health check

The RSSB health check reviewed the completed driver assessments on TRP1 and TRP2 scores to determine if there were

significant differences between candidates according to gender, ethnicity and age. The results are shown in Table 61.

Table 61 - Descriptive statistics for TRP1 and TRP2 scores by fairness groups

			TRP scores			
			TRP1		TRP2	
Protected characteristics		*N*	*Mean score*	*SD*	*Mean score*	*SD*
Gender	*Males*	1394	13.85	2.66	23.50	7.89
	Females	92	13.84	2.76	23.32	8.29
Ethnicity	*White*	1230	14.26	2.62	24.26	7.57
	Other ethnicity	252	11.88	3.14	19.74	8.52
Age	*Up to 50 years*	1396	13.83	2.84	23.66	7.89
	Over 51 years	76	14.46	2.39	20.41	8.10
Overall sample		1486	13.85	2.86	23.49	14.88

There were no significant differences on the TRP scores according to gender or ethnicity. However, there were significant differences in the TRP2 score according to age ($t(1472) = 3.50$, $p = <.001$, where participants aged 21 to 50 scored significantly better than older candidates aged 51+.

The RSSB health check found some differences in pass rates between different groups.

The results for TRP1 are presented in Table 62. This shows that the current pass rates were significantly lower for black candidates than for white British candidates (Chi-square = 91.711, df = 1, $p < 0.001$) and this was less than four-fifths of the pass rate for the majority group.

The pass rates were also significantly lower for Asian candidates compared with white British candidates (Chi-square = 17.575, df = 1, $p < 0.001$) but this difference was not less than four-fifths of the pass rate for the majority group.

Table 62 - Pass rates for the TRP part 1 according to ethnic group

Ethnic group	Pass (%)	Fail (%)	No. assessments in analysis
Black or black British	71	30	132
Asian or Asian British	83	18	80
White – British	94	6	1186
Totals	91	9	1398

The results for TRP2 were very similar to the findings for TRP1. These findings are presented in Table 63. The current pass rates were significantly lower for black candidates than for white British candidates (Chi-square = 38.171, df = 1, $p < 0.001$). The pass rates for black candidates was marginally less than 4/5ths of the pass rate for white British candidates.

The pass rates were also significantly lower for Asian candidates than for white British candidates (Chi-square = 12.637, df = 1, $p < 0.001$) but did not fail the four-fifths rule.

Table 63 - Pass rates for the TRP part 2 according to ethnic group

Ethnic group	Pass (%)	Fail (%)	No. assessments in analysis
Black or black British	72	28	132
Asian or Asian British	78	23	80
White - British	90	10	1189
Totals	88	12	1401

These results suggest that there are fairness concerns with the current cut offs for the TRP1 and TRP2 and these issues should be reviewed to reduce the potential adverse impact.

The review of OPC summary evidence also found significant ethnic group differences by about two-thirds of a standard deviation. This seems to affect black candidates more than Asian candidates. There were no significant group differences relating to age or gender. The data from T948 was not analysed due to difficulty identifying which version of the TRP candidates had undertaken previously, and as much, it was not possible to determine which pass mark applied.

6.3.3.9 Administration time and cost

TRP1 requires the use of a test booklet and a tape / CD player. The information to be learned is played to the applicants while they read it. Candidates then have a further five minutes to read the information and take notes before the information sheet is removed.

There is a further lapse of two minutes before a question and answer booklet is handed out. Multiple choice questions are asked based on the information that has been read and learnt. The time to answer questions is seven minutes.

TRP2 uses an instruction and question and answer booklet which requires applicants to follow lengthy instructions about a set of rules and then attempt to apply these rules to answer the 43 multiple choice questions. The test is relatively easy to administer and easy to score. The time to answer questions is eight minutes.

The total administration time takes approximately 45 minutes which includes moving from part one to part two.

6.3.3.10 Overall conclusion about the method

The T340 review (RSSB, 2005) recommended the retention of the TRP1 and TRP2. These existing assessment methods were considered a reasonable match for the mandatory selection criteria of memory and reasoning so no alternatives were considered during this research.

Both parts of the TRP have shown more consistent criterion related validation evidence than the other tests in the current assessment centre process. There is consistent evidence of the predictive validity of the tests with training related performance measures. There is also evidence that the tests have acceptable levels of reliability, content validity and face validity. The relevance of the tasks to real train driver work is clear and their appropriateness recognised.

There is evidence that different ethnic groups perform statistically differently in this test but this has been reviewed and the proposed new scoring rules would reduce the difference in pass rates between different groups (see Section 7). In conclusion, TRP1 and TRP2 have good psychometric properties and are recommended for future use in the driver psychometric assessment process.

6.3.4 CBI

6.3.4.1 Overview

The Criterion Based Interview (CBI) is a semi-structured interview which is focused on applicants' past experiences. Each section of the interview starts from a description given by applicants – in a candidate interview form (CIF) - of what they consider to be a relevant example for the selection criteria the CBI is designed to cover. After the opening question, interviewers have some freedom to ask follow-up questions. The interview covers six selection criteria:

- Follows set rules and procedures
- Conscientiously works to meet training and job demands
- Remain calm in emergency and stressful situations
- Proactive and tenacious

- Can spend time alone and does so effectively
- Ability to communicate effectively verbally and in writing

Scores for each of these criteria were evaluated in the analysis of the CBI.

Interviewers are trained in the use of a scoring system based on behavioural indicators. Interview performance on each criterion is graded from A to D but only A and B are passes.

6.3.4.2 Sample

T340 review

Data were collected from the industry database for assessments between 2000 and 2003, and analyses are based on the 35-255 applicants for whom CBI data were available. It should be noted that these analyses were limited by the small amount of data available. The CBI criteria 'Proactive and tenacious' and 'Can spend time alone and does so effectively' were only introduced to the assessment centre process in the year before the project started and CBI scores on these selection criteria were only available for 35 drivers for whom there was also performance data.

RSSB Health check

The analysis includes downloaded data (from the RACF industry database provided by Springfield Training) relating to 1,968 assessments conducted between 1st April 2010 and 31st March 2011.

Review of OPC summary evidence (2012)

Four OPC studies are reported where the CBI has been evaluated. In three of these studies the exact sample varies for individual calculations due to missing data. Information on sample sizes is provided in Table 64.

Table 64 - sample sizes from OPC studies

Study date	Min. sample size	Max sample size
2002	19	19
2007	17	66
2009	36	154
2011	100	132

T948 trials

CBI data were successfully obtained for 162 participants. Of these participants, manager ratings of behaviour were available for 52 participants, manager ratings of communication for 82, operational performance data were available for 62, and training performance data were available for 25 participants.

6.3.4.3 Criterion validity

T340 review

This review concludes that there is little evidence for the validity of the CBI. Of the data in this analysis, only two of the criteria assessed using the CBI ('Follows set rules and procedures' and 'Proactive and tenacious') showed any significant concurrent relationships, the former producing two low but statistically significant (p= <.05) correlations of .13 with safe performance and classroom training, the latter correlating.35 with procedure-based work (r(33)=.35, p= <.01). Otherwise, there was no consistent evidence for the validity of other parts of the CBI although some evidence has been reported in previous studies.

It should be noted that these analyses were limited by the small amount of data available (N =35 for the criteria 'Proactive and tenacious' and 'Can spend time alone and does so effectively').

Review of OPC summary evidence (2012)

Table 65 summarises validation evidence for the CBI from studies conducted by OPC in 2002, 2007, 2009 and 2011. The average validities have been weighted by sample size (ie studies with smaller samples have less weight than those with a larger sample).

Table 65 - Criterion validity summary for the CBI scales

CBI Scores	Training: Average Validity	Performance: Average Validity
Follows set rules and procedures	.01	.08
Conscientiously works to meet training and job demands	.07	.05
Remain calm in emergency and stressful situations	.00	.00
Proactive and tenacious	.00	.03
Can spend time alone and does so effectively	.00	.03
Ability to communicate effectively verbally and in writing	.08	.04

There was no evidence supporting the section of the interview which covers 'Remain calm in emergency and stressful situations'. For all the other sections there was at least one study where a good correlation has been observed. However, across studies the findings are very inconsistent with many studies showing no significant relationships for most of the interview sections.

The validities of three parts of the CBI ('Follows set rules and procedures', 'Conscientiously works to meet training and job demands' and 'Ability to communicate effectively verbally and in writing') showed some evidence of generalisability across different studies but the average validities were low. There was very little evidence to support the value of the assessments of 'Proactive and tenacious' or 'Can spend time alone and does so effectively', and none at all for the assessment of 'Remain calm in emergency and stressful situations'.

T948 trials

Using data from the RACF database, analyses were run to examine the correlation between the six CBI scores and job performance measures.

Table 162 in appendix G provides details of all the correlations between CBI scores and performance data. Of the 29 relationships predicted between CBI scale scores and operational and training performance, no significant results were found in the expected direction. Some significant results were found in the wrong direction.

6.3.4.4 Content validity

The CBI is designed to measure the current set of behavioural selection criteria. The current set of criteria is different to those that have more recently been accepted by industry. The differences are summarised in Table 66 below:

Table 66 - Comparison between CBI criteria and updated behavioural criteria

New behavioural criteria	Current behavioural criteria measured by CBI	To note
Conscientiousness (dependability, attitude to work and people, commitment to work, attention to detail, ability to check and not make assumptions, rule compliance)	Conscientiously works to meet training and job demands	The CBI broadly covers the new behavioural criteria but is not specifically designed to cover each of the sub-criteria.
	Follows set rules and procedures	
DCS (calmness under pressure, reactivity to stress, proactive, tenacious, assertiveness)	Remain calm in emergency and stressful situations	Assertiveness is not measured by the CBI.
	Proactive and tenacious	
TLS (sociability, sensation seeking need for stimulation).	Can spend time alone and does so effectively	The CBI broadly covers the new behavioural criteria but is not specifically designed to cover each of the sub-criteria.
Communication (spoken, written, reading, listening)	Able to communicate effectively verbally and in writing	The written aspect of the communication criterion is assessed in a basic qualitative way using the CIF. Only verbal communication (speaking and listening) is assessed as part of the MMI. Written communication is measured using the WCT.

6.3.4.5 Construct validity

Convergent validity

T340 review

Correlations between CBI scale scores and managers' ratings of drivers' attitudes to work (produced from an average of training data or formal assessment data where the former was not available) did not demonstrate good convergent validity. Correlations were expected to be positive but four of the six criteria showed negative correlations, and no correlations were significant (Follows set rules and procedures r(254) = -.12, Conscientiously works to meet training and job demands r(254) = .05, Remain calm in emergency and stressful situations r(254) = .05, , Proactive and tenacious r(35) = -.04 and Can spend time alone and do so effectively (r(35) = -.03, Ability to communicate effectively verbally and in writing r(254) = -.15).

T948 trials

Analyses were run to explore whether the CBI main scale scores correlated significantly in the expected direction with manager ratings of the original selection criteria (ie in line with main CBI scales).

No significant correlations were found between the CBI main scale scores and the related manager ratings of behaviour. However, the CBI measure of 'Conscientiously works to meet training and job demands' and the manager rating of 'Follows set rules and procedures' was significantly correlated in the expected direction $r(50) = .26$, $p = .03$.

Of more relevance are the Pearson correlations calculated to determine the strength of correlation between the behavioural criteria as measured by the CBI, and manager ratings of the updated behavioural selection criteria. The results are provided in Table 67.

Table 67 - Pearson correlations between CBI scales and manager ratings of behaviour (updated behavioural selection criteria)

Manager ratings of behaviour		CBI Scores				
		Follows set rules and procedures	*Conscientiously works to meet training and job demands*	*Proactive and tenacious*	*Remain calm in emergency and stressful situations*	*Can spend time alone and does so effectively*
Consc.	*r*	.01	.06			
	P	.47	.33			
	N	52	52			
DCS	*r*			-.12	.08	
	P			.23	.30	
	N			40	52	
TLS	*r*					.01
	P					.48
	N					40

No significant correlations were found in the expected direction between the CBI main scale scores and the managers ratings of the updated behavioural criteria.

Verbal communication

Pearson correlations were calculated to determine the strength of correlation between communication as measured by the CBI (speaking, listening and written), and the manager's ratings of overall communication (speaking, listening, written and reading). The results are displayed in Table 68.

Table 68 - Pearson correlations between CBI communication score and managers' ratings of communication

Manager ratings of communication		*CBI Ability to communicate effectively verbally and in writing score*
Consolidated overall communication	*r*	-.07
	p	.27
	N	82
Verbal communication	*r*	**.27**
	p	.03
	N	52
Listening	*r*	**.23**
	p	.05
	N	52
Overall written communication	*r*	**.30**
	p	.01
	N	52
Reading	*r*	**.28**
	p	.02
	N	52

The analysis showed that there was a significant relationship between communication as measured by the CBI, and managers' ratings of verbal communication, listening, written communication and reading. However, it did not correlate significantly with a consolidated overall communication measurement which brought together manager ratings from the current trial with manager ratings from the previous trials. This suggests that the finding may not be stable with a larger sample.

6.3.4.6 Discriminant validity

T340 review

The T340 review found that the CBI differentiates better between the various selection criteria than is often the case with interviews. The correlations between the ratings on the six criteria were moderate (averaging r = .32). Furthermore, the CBI seems to provide sufficiently different information to the tests in the driver selection process, the correlations again typically being in the range .3 to .35.

In this study it was found that the pass rate for the interview process were very high - 95%. Reasons for this could include the finding that in some cases applicants had already been selected on the selection criteria addressed by the CBI during the shortlisting and sifting stage. A very high pass rate could indicate that the assessment method does not add much useful information to the assessment. However, the more recent health check analysis found a lower overall pass rate of 68% based on a total of 596 assessments. This suggests the CBI is being used to distinguish between candidates.

T948 trials

The correlations between the ratings on the six criteria were moderate (averaging r = .37). The number of significant correlations observed between ratings on the CBI and other assessment method scores did not exceed the number that would be expected by chance. Please refer to Appendix H for the full results.

These results suggest that the CBI has good discriminant validity.

6.3.4.7 Face validity

T340 review

Concerns were expressed by driver managers about the value of the CBI and particularly the selection criterion ‘Remain calm in emergency and stressful situations’ ratings. The CBI process requires applicants to identify an emergency or stressful situation that they have had to deal with. If the applicant cannot identify such a situation, this is sometimes counted as an automatic fail and certainly reduces the chances of a high rating. However, this is not a valid conclusion. It is not possible to say that someone will not be able to handle a stressful situation in the future simply because they have not experienced one in the past.

In the T340 report it is recommended that this selection criterion should be assessed by some other means in the assessment centre. The same argument can be applied to the other criteria covered by the CBI.

6.3.4.8 Reliability

It has not been possible to access any OPC information about the reliability of the CBI in any of the reviews. However, it is worth noting that interviewers are trained in both the questioning and scoring process and are re-trained on a bi-annual basis. One of the purposes of the training and re-training is to increase interviewer reliability.

In the T948 project, further to the comments recorded in the T340 review, assessors expressed concerns about the reliability of the CBI because it is not a standardised measure of written communication (candidates write about their own experience, rather than all writing about the same thing). Assessors felt that the reliability of the measurement of written communication should be improved.

6.3.4.9 Fairness

T340 review

Due to the restriction within the CBI to past experience, the interview could be discriminating against less experienced high-potential candidates. Chi-squared tests comparing the numbers of candidates with and without rail experience being awarded an A, B or fail on the CBI demonstrated that experienced candidates were significantly more likely to be awarded an A grade than less experienced candidates on all criteria apart from Proactive and tenacious and Can spend time alone and does so effectively ([X2 (1, N = 591 to 593) = 9.29 to 32.37, p <.01 to p= <.001].

Health check review

There was no significant difference in the pass rates of males and females, or according to age or ethnicity.

Review of OPC summary evidence (2012)

Of the studies submitted by OPC for the review, there does not appear to be any information with regard to adverse impact.

T948 trials

No significant differences were found on the CBI scores according to gender, age or ethnicity (N = 133 for Proactive and tenacious and Can spend time alone and does so effectively and N = 158-161 for all scales).

6.3.4.10 Administration time and cost

The CBI is comparable to other interviewing processes in its ease of delivery.

One of the reasons provided by the T948 steering group for dissatisfaction with the CBI is that in some instances interviews could take as long as 90 minutes. On average however, the duration

of CBI administration is approximately 45 minutes (plus 25 minutes for completion of the CIF, and 5-10 minutes for the interviewer to familiarise him / herself with the CIF).

According to information collected from members of the T948 steering group in 2012, the following costs are associated with administering the CBI:

- Initial training costs of approximately £2500 per delegate.
- Interviewer reassessment process every two years which if failed costs approximately £440.
- Materials cost £4.85 per candidate.

Concerns have been raised by assessment centre staff about difficulties in arranging for re-training sessions which has implications for the staffing and running of the assessment centres.

6.3.4.11 Overall conclusions about the psychometric assessment method

The results of the various reviews of the CBI suggest that there is currently limited evidence for the validity of the CBI. Some fair sized correlations have been achieved on all of the scales apart from 'remain calm in emergency', but these results do not appear to be consistent.

Analyses of the construct validity of the CBI communication scale have produced mixed results.

The CBI does not cover all aspects important to the driver role as identified by subject matter experts involved in the T340 review of the selection criteria. There is no measure of assertiveness, and the measure of written communication has been identified by assessors as needing some improvement. Other recommendations that have been made for improvement include the need to incorporate situational questions to measure potential, to complement the interview with another measure of the criteria, and to use a rating scale with more levels of pass (providing they allow meaningful distinctions to be made).

According to the T340 review the CBI does provide valuable information beyond the cognitive tests, and there is some support for the different scale scores within the CBI. However, there is very limited information about the reliability of the CBI.

There is no evidence to suggest that the CBI discriminates by age, gender or ethnicity, but it is likely to put people with less experience at a disadvantage.

7 Development of overall psychometric assessment process and scoring rules

7.1 Introduction

This section details how the recommended psychometric assessment process was designed taking into account all the findings from T340 (RSSB, 2005), T628 (RSSB, 2010) and T948. The assessment process was designed to cover each of the selection criteria using the most suitable assessment method(s) from those that have been evaluated. The section describes, in detail, the rationale for what assessment methods are recommended to assess each selection criterion and what the associated scoring rules are. The recommended scoring rules would be used to make a decision about whether each candidate is eligible to be considered for a job as a train driver.

The overall objectives of the scoring rules are to:

- Set pass marks to screen out potentially unsafe candidates. The pass marks should be designed to assess candidates against a minimum standard where scores below the pass mark indicate that the candidate might not have sufficient aptitude to perform safely. They should not be set to identify potentially exceptional candidates.
- Produce a reasonable overall pass rate for each assessment method. The proportion of candidates that would be screened out by each assessment method should be reasonable given the accuracy with which each assessment method can distinguish between good and poor candidates. The overall pass rate is currently 38% and train operating companies seem satisfied with this rate, and so the revised process should produce approximately the same pass rate. The practicalities of recruitment are such that a certain percentage of candidates need to pass in order to have enough candidates to choose from for the roles available. Too many or too few candidates passing the process would cause problems. However, the need to set a pass mark at a safe level should take priority over this consideration.
- Where more than one method or score is shown to be a reliable and valid indicator of a selection criterion, combine those scores effectively to produce an overall score for each selection criterion. This approach conforms to good practice in using more than one measure to assess a selection criterion (British Psychological Society, 2006).
- Introduce bands of pass where these allow for meaningful differentiation between candidates.

Table 69 shows which assessment methods were considered for the assessment of each selection criterion:

Table 69 - Assessment methods considered for each selection criterion

Selection criterion	Assessment methods considered
Attention	Group Bourdon TEA-Occ SIMKAP
Vigilance	WAFV VIGIL
Memory	TRP1
Reasoning	TRP2 SIMKAP
Perception	TAVTMB ZBA
Reaction time	WAFV VIGIL TAVTMB
Hand coordination	2HAND
Conscientiousness	Situational Judgement Exercise Multi-Modal Interview CBI
Dealing with challenging situations	
Tolerance for low stimulation	
Communication	Multi-Modal Interview Written Communication Test CBI

7.2 Method for the development of the process and scoring rules

The information from the evaluations in T340, T628 and T948 was used to decide what assessment methods and particular scores were the most suitable to be recommended. The scoring rules were then defined to provide a framework for how the assessment methods would be used together to make a decision about whether a candidate meets the minimum standard or not.

Several approaches were used to determine the scoring rules for each assessment method score. The process was iterative; if the results of the first approach did not provide clear information to use for scoring, the analysis would progress to the next approach in the

list and so on. In most cases, the final recommended scoring rules took into account evidence from several different types of analysis. The approach taken for each assessment method was as follows:

- Assessment method scores and the related job performance scores were split into percentiles and expectancy tables were created. Expectancy tables were useful as they display information about the proportions of people with perfect records and varying degrees of incidents at different test score intervals. They were examined to determine whether there was a clear point in the range of assessment method scores below which the majority of poor job performance scores sit, and above which the majority of good job performance scores sit. This would indicate a suitable pass / fail cut-off. This approach did not often indicate a clear cut-off due to restricted range in job performance (ie no or very few participants had poor job performance scores).
- Regression techniques were used to model what assessment measure score a poor job performance score is likely to be associated with. If the associated assessment method score was within a plausible range, this would indicate a suitable cut-off.
- Evidence from the test manuals was consulted to look for guidance relating to the interpretation of the scores.
- The lowest assessment method scores obtained by trial participants were considered as potential cut-offs, taking into account the job performance scores obtained by the lowest scoring participant(s). This was based on the assumption that participants in the trial consisted mainly of drivers and trainees with reasonable performance and was therefore the equivalent of acceptable candidates. For this approach to be used, the lowest score needed to be at a reasonable level ie not too high or too low compared to how the candidate pool is likely to perform. When the lowest assessment method scores were associated with poor job performance (eg negative evidence in the interview such as admitting to breaking rules at work), meaningful scoring decisions were set around the proportion of positive and negative evidence.
- If none of the above approaches produced clear information from which to determine scoring decisions, the cut-offs were set around percentile ranks so that a suitable proportion of candidates were filtered out of the process.

Once scoring rules for individual assessment method scores were determined, a similar process was used to establish how to effectively combine scores to produce a final pass/fail cut-off for each selection criterion. Each assessment method score

contributing to an overall selection criterion score was equally weighted unless the validity evidence suggested that one score should carry more weight than another.

Before finalising the scoring rules, they were applied to the trial sample to see how many trial participants would pass or fail. Qualified drivers and trainees represent a group with acceptable levels of aptitude so if a large proportion of them would fail using the new scoring rules then the scoring rules were set too high. The pass rates of different demographic groups were checked to assess fairness. The pass rate of the candidate population was predicted to assess whether the overall process would be too lenient or too strict.

7.3 Development of the assessment process for cognitive and psychomotor criteria

7.3.1 Introduction

This section details the approach taken to build the recommended assessment process for the cognitive and psychomotor criteria. It explains how the process was formulated in consideration of the assessment methods to assess each selection criteria, the particular scores to use and how these scores would be used to make pass and fail decisions. In addition, the various assessment methods were considered in terms of how they would work together as an overall process.

7.3.2 Building the overall assessment process

The recommended psychometric assessment process and assessment of each selection criteria were designed taking into account all the available evidence on the assessment methods. This included the RSSB study trials T340, T628 and T948, test manuals and other evaluations of potential psychometric assessment methods that were made available to RSSB in sufficient detail.

The recommendation of which assessment methods to use was led by consideration of which particular scores to use. In order for an assessment score to be considered, it needed to demonstrate acceptable levels of content validity, construct validity, criterion validity, reliability and fairness. In order to determine the scores for each selection criterion, each potential score was compared based on these criteria.

Following the review of the available test scores, particular scores were selected for each selection criterion based on their individual contribution. The results are described in the following sections.

7.3.2.1 Attention

The selection criterion for attention is split into two sub-criteria: selective attention and divided attention. Selective attention is defined as the ability to differentiate between different sources of information and attend selectively to them. Divided attention is defined as the ability to switch attention between different sources of

information. Assessment methods were therefore considered in light of their ability to measure selective and / or divided attention.

Three assessment methods were initially considered for the assessment of attention. These methods were TEA-Occ, SIMKAP (both trialled as part of T628) and the Group Bourdon (one of the current assessment methods). A full description of TEA-Occ and SIMKAP is available in the published T628 report (RSSB, 2010).

Scores were short-listed from these three assessment methods based on evidence of their content validity, construct validity and reliability which were all considered acceptable. The assessment scores were then listed and compared in terms of their criterion validity and the extent they could be considered fair based on the available data. The results are summarised in Table 70.

Table 70 – Evaluation of assessment scores as a measure of attention

Selection criteria	Assessment method	Score	Criterion validity		Fairness
			Average validity	*Largest observed correlation*	
Selective attention	Group Bourdon	Total production	0.03 to 0.25 (training), 0.04 to 0.14 (performance)	0.29 (training), 0.32 (performance)	The pass rate for black and Asian candidates is less than 4/5ths of the pass rate for White British candidates.
		Total faults	0.02 to 0.14 (training), 0.00 to 0.07 (performance)	0.38 (training)	
		Total omissions	0.03 to 0.26 (performance), 0.01 to 0.14 (training)	0.41 (performance)	
	TEA-Occ	Visual - Map search (1 min and 2 min)	No significant correlations with training or performance. Overall the results don't seem to be significantly different from chance.		Trial data and test manual suggest over 50s do less well (T628 report).
		Visual - Telephone search (number of circles and time per target)	No significant correlations with training or performance.		Trial data and test manual suggest over 50s do less well (T628 report).
		Auditory - Lift counting with distraction	0.20 (performance), 0.37 (training)	0.69 (performance)	Mixed evidence for age - older people may do slightly better or slightly worse.
	SIMKAP	Speed of perception baseline	No significant results with performance, 0.50 (training). Analysis suggests number of	0.51 (training)	Over 50s score lower than younger people.

			significant correlations are same as expected by chance.		
		Accuracy baseline	No significant results with performance, 0.50 (training)	0.50 (training)	
Divided attention	TEA-Occ	Tel. search with counting - time per target score	0.20 (performance)	0.25 (performance)	Over 50s score lower than younger people. No significant differences according to gender or ethnic group.
		Tel search with counting – dual task decrement	0.22 (performance), no significant results with training.	0.32 (performance)	
	SIMKAP	Simultaneous capacity	0.19 (performance), no significant results with training.	0.19 (performance)	Over 50s score lower. No difference for gender and ethnic group.
		Accuracy simultaneous	0.10 (performance), some significant results with training but in the wrong direction.	0.37 (performance)	
		Speed of perception simultaneous	0.18 (performance), 0.24 (training)	0.24 (training)	
		Mixed questions simultaneous	0.44 (performance), no significant results with training.	0.44 (performance)	

Selective attention

Firstly, considering the Group Bourdon, the total production score and total omissions score are recommended to assess selective attention. The evidence of their relationship with training and operational performance is acceptable although some studies have shown inconsistent results. The total faults score is not recommended as a review of multiple studies provided by OPC suggest that there is little if any evidence of criterion validity for this particular score. There were some ethnic group differences within the Group Bourdon scores which provides a case for the pass marks to be adjusted to improve the fairness of this test. This is discussed in Section 6.3.3.8.

The lift counting with distraction task of the TEA-Occ was significantly correlated with both training outcomes and operational driving performance. This score is recommended. However, there were no significant correlations between the map search and telephone search subtests with either training or operational performance data. There were also differences related to age where older candidate performed less well. Only the lift counting with distraction subtest will be recommended as a measure of selective attention.

The results for SIMKAP were weak and whilst the perception accuracy baseline scores were significantly related to training outcomes, further analysis suggests that the correlations observed are no stronger than would be expected by chance. In addition, the correlations observed in the T628 trials could not be replicated with additional performance data in the T928 trials which suggests that these correlations are unstable. The SIMKAP scores are not recommended for these reasons.

In summary, it is recommended that selective attention will be assessed by the Group Bourdon total production and total omissions scores and the TEA-Occ lift counting with distraction score. Whilst the criterion validity of both these methods is moderate, using both methods in combination is considered to provide a more robust measure of selective attention, particularly as visual attention is assessed by the Group Bourdon and auditory attention is assessed by the TEA-Occ.

Divided attention

To assess divided attention, two scores from the TEA-Occ and four scores from SIMKAP were considered.

For TEA-Occ, the criterion validity of the dual task decrement score is stronger than the time per target score so only the former is recommended.

For SIMKAP, whilst there were some significant correlations with training and performance, these correlations were sometimes in the wrong direction and the strength of the correlations were no better

than would be expected by chance. The SIMKAP scores are not recommended.

In summary, the dual task decrement score in the TEA-Occ is recommended to assess divided attention.

7.3.2.2 Vigilance

During the T628 trials, the lottery subtest of the TEA-Occ was evaluated to determine its appropriateness as a measure of vigilance. Unfortunately, the lottery subtest demonstrated poor levels of criterion validity and there were no significant correlations with safety performance. Several vigilance tests were considered and two were shortlisted to be evaluated as part of the T948 trials to close this gap. These tests were WAFV and VIGIL. Further information about how these tests were shortlisted is detailed in Appendix A.

The results of the T948 trials demonstrated that VIGIL did not have acceptable levels of criterion validity so only the WAFV scores were considered in the scoring rules. The results of the WAFV scores are summarised in Table 71.

Table 71 - Evaluation of assessment scores as a measure of vigilance

Selection criteria	Assessment method	Score	Criterion validity		Fairness
			Average validity	*Largest observed correlation*	
Vigilance	WAFV	Number of missed reactions	-0.20 (performance)	-0.21 (performance)	No significant differences according to gender, ethnic group or age
		Number of false alarms	-0.26 (performance)	-0.36 (performance)	

Both the number of missed reactions and the number of false alarms are recommended to assess vigilance. They both have strong evidence of criterion validity with relevant operational driving performance measures, including SPAD records and train handling. Also, there were no significant group differences with regard to gender, age or ethnicity.

7.3.2.3 Memory

Memory is defined as the ability to learn, recall and apply job related information in appropriate time limits. The TRP part one was selected as it measures the ability to learn fictitious rules information, remember it and apply it in order to answer a series of questions.

The TRP part one is in use already as part of the current assessment process. The results of the TRP part one evaluation are summarised in Table 72.

Table 72 - Evaluation of assessment scores as a measure of memory

Selection criteria	Assessment method	Score	Criterion validity		Fairness
			Average validity	*Largest observed correlation*	
Memory	TRP	TRP1	0.31 (training), no significant correlations with performance	0.41 (training)	The pass rate for black and Asian candidates is less than 4/5ths of the pass rate for white candidates.

The TRP1 score is recommended to assess memory. The results confirm that the TRP1 has a moderate to strong relationship with training, but not with performance. There are some ethnic group differences with the current pass marks which presents a case for these pass marks to be adjusted. This is discussed in section 6.3.3.8.

7.3.2.4 Reasoning

Reasoning is defined as the ability to solve problems and make decisions and relates to train driving tasks such as fault diagnosis and interpreting information from instrumentation. There were two potential assessment methods for measuring reasoning: TRP part two and SIMKAP. The results of these scores are summarised in Table 73.

Table 73 - Evaluation of assessment scores as a measure of reasoning

Selection criteria	Assessment method	Score	Criterion validity		Fairness
			Average validity	*Largest observed correlation*	
Reasoning	SIMKAP	Stress tolerance	0.03 (performance)	0.23 (performance)	No significant differences for age, gender or ethnic group.
	TRP	TRP2	0.20 (performance), 0.16 (training)	0.41 (training)	Significant differences for black and Asian candidates.

The evidence of criterion validity for SIMKAP was weak as the average correlation with performance data was only r = .03. Some of the expected correlations with performance data were also in the

wrong direction so these results are considered as unstable. In addition the construct validity of stress tolerance as a measure of reasoning is less obvious. Therefore the SIMKAP stress tolerance score was discarded from consideration.

However, the TRP2 was considered as a suitable assessment of reasoning because it requires the candidate to apply information they have learnt in order to make decisions. The evidence of criterion validity is more favourable for TRP2 with the average correlation meeting acceptable levels. The TRP2 is in use already as part of the current assessment process and this assessment method is already accepted by the industry. Like TRP1, there were some ethnic group differences with the current pass marks which are discussed in section 6.3.3.8

7.3.2.5 Perception

Two assessment methods were considered for the assessment of perception – the Tachistoscopic Traffic Test (TAVTMB) and the Time Movement Anticipation Test (ZBA) – both of which were trialled as part of the T628 study. Further information can be found about these tests in the published T628 report (RSSB, 2010). The results of these scores are summarised in Table 74.

Table 74 - Evaluation of assessment scores as a measure of perception

Selection criteria	Assessment method	Score	Criterion validity		Fairness
			Average validity	*Largest observed correlation*	
Perception	TAVTMB	Overview	0.26 (performance), 0.78 (training)	0.78 (training)	Significant differences between ethnic groups but only 0.4 of a standard deviation so should not break the four-fifths rule with a 90% pass rate.
	ZBA	Median direction deviation (total)	T628 analysis: 0.54 (performance), -0.19 (training). T948 analysis: no significant correlations found.	0.57 (performance)	ZBA manual states 'nothing to indicate that ZBA is contrary to fairness' but evidence of the analysis is not presented. Large differences between whites and others on this score but small sample with outliers. No difference for age or gender.
		Median time deviation (total)	T628 analysis: no significant correlations with performance, 0.32 (training). T948 analysis: no significant correlations found.	0.32 (training)	

There was evidence that the TAVTMB overview score has strong criterion validity with training outcomes and moderate criterion validity with operational driving performance, including safe performance, procedure based work and train handling. The number of correlations found was significantly more than would be expected by chance.

In comparison, the number of correlations between the ZBA scores and criterion data were not significantly stronger than would be expected by chance. Whilst the T628 study reported significant correlations with safe performance, train handling and procedure based work, these correlations were corrected for range restriction and may have been inflated. Additional performance data were collected as part of the T928 trials and these results were not replicated. The ZBA scores are therefore not recommended. This means that a specific assessment of anticipating speed and

distance is not included in the current assessment process. It is recommended to take the assessment of perception forward using just the TAVTMB overview score and if it emerges that there is a significant gap in trainee driver skills in this area then an additional assessment method can be trialled and introduced at a later date.

7.3.2.6 Reaction time

Reaction time is defined as a quick and adequate response to simple and complex visual and acoustic stimuli and the associated quality of performance. Three assessment methods were considered for the assessment of this criterion – time scores from WAFV, TAVTMB and DTG. The results of these scores are summarised in Table 75.

Table 75 - Evaluation of assessment scores as a measure of reaction time

Selection criteria	Assessment method	Score	Criterion validity		Fairness
			Average validity	*Largest observed correlation*	
Reaction time	WAFV	Mean reaction time	-0.19 (performance)	-0.19 (performance)	No significant differences were found between groups based on age, gender or ethnicity.
	TAVTMB	Working time	T628: no significant correlations found. T928 analysis: -0.19 (performance)	-0.19 (performance)	The difference between the over 50s and the rest are in the range 0.5 - 0.6 of a standard deviation and so might be close to breaking the four-fifths rule with a 85% - 90% pass rate. No evidence of problem with gender or ethnic group.
	DTG	Part 3 good	Review of OPC evidence: 0.08 (performance), 0.07 (training). T948 analysis: no significant correlations	0.31 (training)	Significant differences found between ethnic groups but these effects are less than the 4/5^{ths} rule.
		Self-paced good	Review of OPC evidence: 0.06 (performance), 0.08 (training). T948 analysis: no significant correlations	0.37 (training)	

The number of significant correlations between WAFV mean reaction time and criterion measures is higher than would be expected by chance.

The results for TAVTMB working time were similar. The number of significant correlations between this score and criterion measures is higher than would be expected by chance.

The results for DTG are weak and no significant correlations were found in the T948 trials. The review of OPC studies did find some evidence of significant correlations but the average correlations were very weak.

The WAFV reaction time and TAVTMB working time scores were originally recommended to assess reaction time to simple and complex stimuli. However, the TAVTMB has been replaced by an adaptive version called the Adaptive Tachistoscopic Traffic Test (ATAVT). Whilst the overview score is equivalent for the linear and adaptive version of the test, the working time score is not equivalent so cannot be used as a measure of reaction time.

In summary, only the WAFV reaction time score is recommended for the assessment of reaction time.

7.3.2.7 Hand coordination

The 2HAND is the only measure of hand coordination recommended for inclusion in the new train driver psychometric assessment process. The results of the relevant scores are summarised in Table 76.

Table 76 - Evaluation of assessment scores as a measure of hand coordination

Selection criteria	Assessment method	Score	Criterion validity		Fairness
			Average validity	*Largest observed correlation*	
Hand coordination	2HAND	Total mean duration	-0.30 (performance)	-0.32 (performance)	Significant differences between ethnic groups.
		Total percent error duration	No significant correlations found.		No significant differences on this score.

There was consistent evidence of the criterion validity of the total mean duration score and the correlations found were statistically higher than would be expected by chance. There were no significant correlations found for the total percent error duration score but this score will be retained with a low pass mark as a measure to counter cheating and identify non-serious test takers.

7.3.3 Setting the scoring rules for each cognitive and psychomotor selection criterion

Having chosen which assessment scores to use to assess each selection criterion, the next step was to set the scoring rules. This involved deciding how the test scores should be combined, and what the pass mark should be for each score using the methods described in Section 7.2 of Annex 3.

The following sections present the recommended scoring rules and rationale for the assessment of each cognitive and psychomotor selection criterion.

7.3.3.1 Attention

Scoring and banding

The recommended assessment scores and raw score pass criteria are shown in Table 77.

Table 77 - Attention criteria – Recommended assessment scores

Selection criteria	Assessment method	Score	Raw score pass criteria
Attention	TEA-Occ	Lift counting with distraction – The number of correctly counted strings of tones.	≥ 6
		Dual task decrement – The extent to which performance is reduced when doing two tasks simultaneously.	≤ 4.44
	Paper Group Bourdon	Total production – The amount of stimuli that are checked within the time limit.	≥ 938
		Total omissions – The number of target stimuli that are missed.	≤ 47

Rationale and supporting evidence

TEA-occ

As noted previously, it is recommended that selective attention should be assessed by the Group Bourdon total production score, total omissions scores and the TEA-Occ lift counting with distraction score. Divided attention should be assessed by the TEA-Occ dual task decrement score. Using scores from both assessment methods is considered to provide a more robust measure of attention, particularly as visual attention is assessed by the Group Bourdon and auditory attention is assessed by the TEA-Occ.

For the attention criterion the objective is to screen out candidates who have a very poor ability to maintain attention. In order to set the raw pass mark criteria for the TEA-Occ, the performance of stroke patients and control subjects was compared from the same age category (Robertson et al, 1994). As is clear in Table 78, there are statistically significant differences between stroke patients and control subjects in the lift counting with distraction task. For the lift counting with distraction subtest, the average score of a stroke patient is 5.65 (out of 10). It was therefore considered necessary to set the pass criteria above this point (rounded up to 6) in order to

exclude scores that are associated with attentional deficits such as those suffered by stroke patients or closed head injury patients.

Table 78 – Stroke verses controls on two TEA subtests

Subtest	50 – 64 Control (n = 26)	Stroke patients (n = 39)	t	P
Lift counting with distraction	8.18 (2.8)	5.65 (3.2)	3.45	<0.001
Telephone search – dual task decrement	2.03 (3.4)	3.77 (9.5)	ns	Ns

In order to assess the percentile rank of this pass criterion, normative tables were consulted in Table 79. This table includes normative data from the TEA-Occ standardisation study where 136 candidates attending train driver assessment centres completed the TEA-occ from nine train operating companies across Great Britain. Of this sample, 58 percent were trainee train drivers, 27 percent currently worked as train drivers and the remaining 15 percent worked in various roles within the train industry. Comparing the TEA-Occ norms with control subjects from the TEA norms, it is clear that the TEA-Occ sample have better levels of attention than those of the general population. It can be seen that a cut off of 6 corresponds to the 10th percentile rank of the TEA-Occ norm. Given that this norm is based on existing rail staff, the proportion of potential candidates excluded at assessment centres is likely to be more than 10%. Therefore, the lift counting with distraction score is recommended to be set at a raw score of 6 or above.

The norm group to be used for the implementation of TEA-Occ will comprise consolidated data from the standardisation study (n=136), the T628 trials (n=130) and potentially norm data generated by Network Rail in signaller recruitment. The sample for this norm will be n=>260. This norm will be replaced with train driver candidate norms once sufficient data has been collected following implementation.

Table 79 – Normative data for two TEA-Occ subtests

Subtest	Percentiles	TEA-Occ 21-51 years	TEA 35-49 years	TEA 50-64 years
Subtest 2: Lift Counting with Distraction *Max = 10*	1	1	-	-
	5	4	2	3
	10	**6**	4	5
	25	8	7	8
	50	10	9	9
	75	10	10	10
	90	10	10	10
	95	10	10	10
	99	10	10	10
Subtest 5: Telephone Search with Distraction *(Dual task decrement)*	1	19.85	-	-
	5	**4.44**	7.0	8.0
	10	2.88	4.5	5.0
	25	1.83	2.5	2.5
	50	0.96	1.0	15
	75	0.65	0.5	0.5
	90	0.37	0.0	0.0
	95	0.06	0.0	0.0
	99	-2.50	-0.5	-0.5

For the telephone search while counting dual task decrement score, a similar approach was taken. Table 78 shows that the mean score for a control subject is 2.03 and the mean is 3.77 for a stroke patient although the stroke patients' performance has a high degree of variability as demonstrated by a standard deviation of 9.5. Taking these averages into account, the 10th percentile rank of the TEA-Occ norm corresponds to a raw score of 2.88. This is considered too strict as the purpose of the cut off is to select out potentially unsafe candidates. Therefore, the 5th percentile was considered which corresponds to a raw score of 4.44. Given that this norm is based on existing rail staff, potential candidates excluded at assessment centres is likely to be at the 5th percentile rank or higher. Therefore, it is recommended that the dual task decrement cut-off score is set at a raw score of 4.44 or higher (higher scores represent worse performance).

As a precautionary check, these cut-offs were applied to the T628 trial data in order to determine if these pass criteria identify existing drivers who have experienced operational incidents, and therefore

may have issues with attentional capacity. When these cut-offs were applied to the T948 trial participants, the people who would fail are mostly previously failed candidates (from previous attempts at assessment centres) and a smaller number of existing drivers who have had operational incidents including station overruns.

Group Bourdon

Evidence in favour of the TEA-Occ was not sufficiently strong to justify completely replacing the Group Bourdon in terms of construct and criterion validity. As the Group Bourdon measures visual selective attention, and measures a slightly different aspect of attention than the TEA-Occ, it was decided to use conjunctive scoring rules where candidates need to reach a minimum level on each individual score.

The current pass marks for the Group Bourdon were used as a starting point for recommending pass marks for the updated process. It was found that the omissions score was set very high and excluded many more candidates (67%) than could be justified on the basis of the validity of this particular score alone. In addition, there was a strong difference in performance on this score between white and black candidates. In fact, the pass rate for black candidate using the current cut offs and based on RACF data from 2010/2011 showed that the pass rate for black candidate was only 52% as opposed to 72% for white candidates. This clearly fails the $4/5^{ths}$ rule.

In order to reduce adverse impact, the pass mark for the omissions score should be changed significantly from 0 – 22 to 0 – 47 to reduce the ethnic group differences. This new cut off would make a marked difference to the pass rates as shown in Table 80. This proposed change should result in pass rates that conform to the 4/5ths rule and should still maintain safety.

Table 80 – Pass rates for Group Bourdon omissions based on current and proposed cut offs

Ethnicity	% *pass rate* with current scoring (pass= 0-22)	Passes 4/5ths rule	% *pass* with proposed scoring (pass=0-47)	Passes 4/5ths rule	N
White	72	Y	93	Y	1242
Other white	60	Y	90	Y	60
Asian	58	Y	86	Y	125
Black	52	N	81	Y	192
			Total N		1619

The current-cut off for the production score of the Group Bourdon does not result in any group differences based on protected characteristics including age, gender or ethnicity. Therefore the production score pass mark is recommended to remain as it is currently and would screen out approximately the bottom 11% of candidates.

The norm group to be used for the implementation of the Group Bourdon will comprise data on train driver candidates from the RACF database between April 2010 and 2011 (n=1681).

Based on this norm group, it is estimated that a maximum of 35% of candidates would be screened out of the candidate pool on the basis of the recommended attention assessment using a combination of the Group Bourdon and TEA-Occ scores. This pass rate is at a similar level to the current process for the assessment of attention.

7.3.3.2 Vigilance

Scoring and banding

The WAFV is recommended to assess vigilance. Two different scores are used as shown in Table 81.

Table 81 - Vigilance criteria – Recommended assessment scores

Selection criteria	Assessment method	Score	Raw score pass criteria
Vigilance	WAFV	Missed reactions – The number of target stimuli that are missed after 1500ms of presentation.	≤ 5
		False alarms – The number of non-target stimuli that are incorrectly responded to.	≤ 8

Rationale and supporting evidence

Vigilance is considered to be one of the most important and critical for safety performance. The missed reactions score is recommended by the test publisher as the primary measure of vigilance. In order to set the cut off for missed reactions, the worst performers from the T948 trials were analysed to determine the relationship between a low test score and any instances of operational incidents. The proposed pass criterion is 0-5 missed reactions which equates to roughly the 10^{th} percentile of the norm group. This pass criterion seems to work well and where trial participants would fail, in the majority of cases their performance has been associated with occasional significant errors related to train handling, Cat A SPADs and station disregards, and regular minor errors including speeding, overruns errors in the preparation of trains. This provides evidence that the pass criterion for the missed reactions score is capable of screening out potential unsafe candidates.

It was decided that the false alarms score should also be retained because otherwise a candidate could pass by responding to *every* stimulus whether target or non-target and this would not be a valid measure of vigilance. The cut off for false alarms is therefore used only to exclude those who are wilfully attempting to cheat the test by pressing the button repeatedly. As the missed reactions and false alarms scores are only moderately correlated ($r = .44$, $p < .001$) it suggests these scores are sufficiently different to assess different aspects of vigilant attention.

As an additional assessment of these proposed scoring rules, a regression using missed reactions and false alarms was undertaken using overall train handling as the measure of performance. This produced significant results ($F = 14.05$, $p < .001$).

For the implementation of the WAFV, a norm sample is recommended which is representative of the general population consisting of 295 individuals (46.4% men, 53.6% women) aged between 16 and 77. The standardisation was carried out between December 2005 and April 2006 under standardised conditions in the research laboratory of the Dr.G.Schuhfried company. This norm will be replaced with train driver candidate norms once sufficient data has been collected following implementation.

The vigilance assessment should be administered following the attention, reasoning and memory assessments. It is predicted that 10% of the remaining candidates would fail the vigilance criteria with the recommended scoring rules.

7.3.3.3 Memory

Scoring and banding

The TRP1 assessment method is recommended to assess memory as shown in Table 82.

Table 82 - Memory criteria – Recommended assessment scores

Selection criteria	Assessment method	Score	Raw score pass criteria
Memory	TRP	TRP1 - number of correctly answered questions	≥ 9

Rationale and supporting evidence

Only one score will be used to assess memory - the TRP part one score. As this score is not highly correlated with any other assessment method score it is considered to provide unique information within the psychometric assessment process.

The current pass mark was used as a starting point. Looking at the pass rates across protected characteristics, it was found that the pass rate for black candidates was statistically lower than the pass rate for white candidates and could be considered to be unfair, especially since the results of this assessment method primarily relate to training outcomes and not to safety performance. Therefore, it is recommended to relax the pass mark by one point. When this new cut off was applied to previous assessment centre data from 2010/2011 this reduced the difference in pass rates as shown in Table 83. Adjusting the cut off by one raw score improves the pass rate from 71% to 81% for black candidates. It was not possible to look at the relationship with of this cut off with performance data as every candidate in the RACF data which was analysed has passed TRP1.

Table 83 - Pass rates for TRP1 based on current and proposed cut offs

Ethnicity	% *pass rate* with current scoring (pass= 10+)	Passes 4/5ths rule	% *pass* with proposed scoring (pass= 9+)	Passes 4/5ths rule	N
White	95	Y	95	Y	1186
Other white	87	Y	87	Y	45
Asian	83	Y	83	Y	80
Black	71	N	81	Y	132
			Total N		1619

For the implementation of TRP1, a norm sample extracted from the RACF database from April 2010 to 2011 (n= 1486) and is representative of the train driver candidate pool is recommended.

Based on this norm sample, it is estimated that a maximum of 6% of candidates could fail the memory assessment. However, TRP1 will be performed following the attention assessment and it is predicted that up to 4% of candidates who passed attention would fail on memory.

7.3.3.4 Reasoning

Scoring and banding

The TRP2 assessment method is recommended for the assessment of reasoning as shown in Table 84.

Table 84 - Reasoning criteria – Recommended assessment scores

Selection criteria	Assessment method	Score	Raw score pass criteria
Reasoning	TRP	TRP2 - number of correctly answered questions.	≥ 13

Rationale and supporting evidence

As the TRP2 score is currently used in assessment centres, the current cut off was used as a starting point in setting the pass mark for use within the new recommended process. As with TRP1, pass rates using the current pass mark are significantly lower for black candidates than for white candidates. It is recommended to relax the pass mark by one point to reduce this difference. Data from previous assessments using RACF data from 2010/2011 show that this improves the disparity so that it is in compliance with the four-fifths rule but it is likely that there will still be a difference in pass rates between blacks and whites.

When the new recommended pass mark was applied to assessment data from 2010/2011 the pass rates were 91% and 77% respectively for White and non-white candidates (see Appendix I). However, given the strong relationship between results on this assessment method and performance on the job (see Section 6.3.3.3), it is not considered justifiable to relax the pass mark further due to safety concerns.

For the implementation of TRP2, a norm sample extracted from the RACF database from April 2010 to 2011 (n= 1486) and representative of the train driver candidate pool is recommended.

The new recommended pass mark is set at a point that would screen out approximately 11% of applicants if administered alone. However, the TRP2 would be administered after the attention and memory assessments so the actually failure rate for reasoning is predicted to be around 3-4%.

7.3.3.5 Perception

Scoring and banding

Perception is recommended to be assessed using the overview score from the ATAVT with a pass mark that is equivalent to six or higher on the TAVTMB. (The TAVTMB was upgraded to the ATAVT so the pass criteria needed to be converted). Norm data from Schuhfried was consulted to identify the ATAVT-person parameter estimate corresponding to a raw score of six on TAVTMB overview. It was recommended that a cut-off value on ATAVT would be -1.5066. The ATAVT assessment method is therefore recommended for the assessment of perception as shown in Table 85.

Table 85 – Perception critera - Recommended assessment scores

Selection criteria	Assessment method	Score	Raw score pass criteria
Perception	ATAVT	Overview score	> -1.5066

Rationale and supporting evidence

The ATAVT overview score provides a suitable measure of the ability to perceive elements within a traffic environment. This score measures the ability to perceive and correctly respond to objects in a traffic environment and is significantly correlated with operational driving performance including isolation of safety systems and procedure based work.

Expectancy tables were created to determine where the cut off should be set for this score but unfortunately these tables did not prove to be useful. Instead, regression analysis was undertaken for this score and performance records for the correct isolation of safety

systems which was significant (F = 5.17, p = .03). This cut off was applied to the trial sample to determine the characteristics of the participants who would fail. Four participants in the trial sample would fail; one was a driver and required reassessment on the job, whilst the other three participants were not existing drivers. The recommended pass mark therefore equates to the lowest scores achieved by people in that sample and is likely to identify people who have perceptual difficulties.

For the implementation of the ATAVT, a norm sample is recommended which is representative of the general population consisting of N=1190 individuals. The sample consists of 574 (48.3%) men and 615 (51.7%) women. The expected distribution was calculated from the data of the Austrian census of 2001 and a census carried out in Germany after reunification. The respondents range in age from 15 to 94. The median age is 41 with a mean age of 42 years 4 months and a standard deviation of 16 years 2 months. The standardisation was carried out in 2007 under standardised conditions in the research laboratory of the Dr.G.Schuhfried company. This norm will be replaced with train driver candidate norms once sufficient data has been collected following implementation.

Although this pass mark is set quite high, candidates will already have been assessed on most of the other criteria so the perception assessment is not expected to exclude too many candidates at the later stages of the assessment process. It is estimated that the perception assessment will exclude 4-5% of candidates who have not been excluded by the previous assessments.

7.3.3.6 Reaction time

Scoring and banding

It is recommended that the WAFV mean reaction time score is used to provide a measure of reaction time as shown in Table 86.

Table 86 - Reaction time criteria – Recommended assessment method

Selection criteria	Assessment method	Score	Raw score pass criteria
Mean reaction time	WAFV	Reaction time – The average time taken to respond to the target stimuli.	< 656

Rationale and supporting evidence

The WAFV reaction time score is recommended for the assessment of reaction time because it provides a measure of simple reaction speed to simple stimuli. The pass mark for reaction time was set taking into account the requirements of the train driving task.

Train drivers need to react within a reasonable amount of time but they do not need to have exceptionally fast reactions. To provide an example, train drivers have either 2 or 2.7 seconds to respond to AWS depending on whether the train is a high speed one or not. A study on driving performance was undertaken as part of research project T148 (RSSB, 2005) which found that the average response time to respond to AWS was 0.6 – 0.9 seconds based on OTMR data downloaded on 277 drivers. Therefore, the pass mark for WAFV reaction time is set to exclude only the worst 5% of performers. Whilst the reaction speed requirements of the driving task are lower than the reaction time specified in the pass criteria, the rationale for setting it at this level is that responding to AWS includes additional physical movement as the driver is required to move their hand. The reaction time for WAFV does not include a similar physical movement so the reaction time would be expected to be quicker.

A regression analysis was undertaken between the WAFV reaction time score using the proposed cut-off which produced a significant relationship with SPAD records ($F = 4.57$, $p = .04$). In addition, this cut off was applied to the T948 trial participants to determine their characteristics. Only one participant would fail based on the recommended cut-off but he also would have failed many of the other tests which suggests that a lack of motivation contributed to the test results.

For the implementation of the WAFV, a norm sample is recommended which is representative of the general population consisting of 295 individuals (46.4% men, 53.6% women) aged between 16 and 77. The standardisation was carried out between December 2005 and April 2006 under standardised conditions in the research laboratory of the Dr.G.Schuhfried company. This norm will be replaced with train driver candidate norms once sufficient data has been collected following implementation.

Based on this norm sample, it is estimated that the reaction time assessment will exclude 3-4% of candidates who have not been excluded by the previous assessments.

7.3.3.7 Hand coordination

Scoring and banding

Hand coordination is recommended to be assessed using the 2HAND assessment method overall mean duration and percent error duration scores as shown in Table 87.

Table 87 - Hand coordination criteria – Recommended assessment method

Selection criteria	Assessment method	Score	Raw score pass criteria
Hand coordination	2HAND	Overall mean duration – The average time taken to complete the tracks	< 52.8
		Percent error duration – The percentage of time spent outside the track	< 16.7

Rationale and supporting evidence

The overall mean duration and percent error duration scores are recommended from the 2HAND as they provide a measure of both speed and accuracy of hand coordination.

Expectancy tables were used to determine that the worst performers on this assessment method were also the participants who had poor driving records which confirms that the method works.

Percent error duration is required in addition to overall mean duration otherwise a candidate could pass the assessment by quickly moving the ball to the other end of the track without attempting to stay within the lines.

As the hand coordination selection criterion is imposed by the Driver Licensing directive and was not identified by GB subject matter experts as a key safety critical aptitude for train driving, the pass mark is set low to exclude only the worst 5% of candidates on both scores. Conjunctive rules are recommended so that candidates are required to have reasonable speed and accuracy in their hand coordination ability. These scoring rules identified a driver in the trial sample who had one station overrun.

For the implementation of the 2HAND, a norm sample is recommended which is representative of a sample of Swedish job seekers (n=209) aged between 18-57 years, 57% of which were male and 43% were female, who were tested within the scope of career counselling and training. This norm will be replaced with train driver candidate norms once sufficient data has been collected following implementation.

The 2HAND is recommended to be the final part of the cognitive and psychomotor assessment. It is predicted that it would exclude only 2% of the remaining candidates.

7.4 Development of the scoring rules for behavioural criteria

7.4.1 Situational judgement exercise

7.4.1.1 Recommended scoring and banding

The SJE scoring is used for two purposes; with the MMI scores to make an overall pass/fail decision, and also to inform the MMI (namely, whether additional situational questions are required for any sub-criteria).

To present the SJE results ready for combination with the MMI results, three main scores are produced – one for each of the main behavioural criteria – and banded into one of two bands as shown in Table 88.

Table 88 - SJE main and sub-criteria interim score banding

Main and sub-criteria banding		**Interim Score bands**	
		Lower limit	*Upper limit*
SJE Conscientiousness	*Low*	none	77.49
	Moderate / good	77.5	none
SJE DCS	*Low*	none	77.49
	Moderate / good	77.5	none
SJE TLS	*Low*	none	77.49
	Moderate / good	77.5	none

SJE sub-criteria scores are used to inform questions asked in the MMI. This is achieved by a) banding SJE sub-criterion scores into 'low', 'moderate' or 'good' based on standardised score bands (using the trial sample as a norm group until further data is collected), and b) calculating the consistency of each sub-criterion score ('weak', 'medium' or 'strong'). Candidates with sub-criteria with a 'low' or 'moderate' score, or 'weak' consistency, will be asked a situational question in addition to the default behavioural question in the relevant MMI topic area. Behavioural questions relate to past experiences and situational questions relate to hypothetical scenarios. Further information regarding the MMI is available in Annex 3 Section 2.7.

7.4.1.2 Rationale and supporting evidence

SJE bands for the overall pass / fail decision

It was decided that the SJE should not be used alone to make a pass/fail decision because it is best practice to use more than one method to measure the criteria. Instead, two simple bands would be used in combination with the MMI scores to make pass/fail decisions.

The SJE expectancy tables did not provide a clear indication for the cut-off scores because of the limited range in job performance scores among the participants, and regression analyses did not produce plausible associated scores. For this reason, the score bands were set in relation to where the drivers and trainees with poor job performance ratings scored on the SJE, and adjusted to suit a candidate population. This adjustment process involved a process of standardising the scores (creating z100 scores) and setting the 'low' band at 2.5 standard deviations below the mean to reflect the position of the lowest score on each main criterion.

The analysis showed that the SJE demonstrated a significant linear relationship with manager ratings of behaviour (see section 5.6.4 on criterion validity), and the participants with the very lowest SJE scores had some low ratings of performance. This lends support to the proposal to set the 'low' SJE bands as detailed in Table 88.

SJE bands for the MMI report

The 'moderate' and 'good' bands are merged into one band for the purpose of determining an overall pass/fail from the combined SJE and MMI scores. This is because at this stage, the validity evidence was not strong enough to justify a distinction between SJE 'moderate' and 'good' in making a final pass/fail decision. Only a distinction between 'low' and 'moderate/good' was supported.

However, the trial evidence did support a distinction between the 'moderate' and 'good' SJE bands for the purpose of informing the MMI. The validity evidence was strong enough to justify using the SJE score band to indicate what areas should be explored more deeply during the MMI (job performance ratings are significantly higher among participants scoring 'good' compared to those scoring 'moderate': Conscientiousness ($t(9) = -2.01$, $p = .08$), DCS ($t(56) = -2.61$, $p = .01$), TLS ($t(56) = -1.99$, $p = .05$)).

The parameters of the consistency bands were based on published guidance on critical values of average deviation (Dunlap, Burke and Smith-Crowe, 2003).

A note on the interim score bands

Conservative band limits have been set to represent the fact that drivers involved in the trial had good performance ratings. The candidate sample was not expected to have the same distribution and so as more data are collected this assumption will be tested, and the score bands adjusted accordingly. The development of a

complete norm group will also allow the final results to be presented as percentiles.

If, with the collection of more data to produce a norm group, there is strong statistical support for differentiating between the 'moderate' and 'good' bands (ie a meaningful difference is shown between the bands), this distinction can be used in the final scoring decision.

Fairness

It was expected that there would be no significant differences in SJE bands according to gender, ethnicity or age. Chi squared tests were performed to check for differences, although it should be noted that due to limited sample sizes the results should be interpreted with caution. The results are shown in Table 89, Table 90 and Table 91.

Table 89 - Descriptive statistics for SJE Conscientiousness score bands by demographic groups

Protected characteristics		**Conscientiousness bands**								
		N	**Low**		**Acceptable / good**		**Acceptable**		**Good**	
		N	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*
Gender	*Males*	60	1	2	59	98	9	15	50	83
	Females	9	0	0	9	100	3	33	6	67
Ethnicity	*White*	61	1	2	60	88	9	15	51	84
	Other	8	0	0	8	100	3	38	5	63
Age	*up to 50 years*	57	1	2	46	98	9	16	47	83
	over 51 years	12	0	0	12	100	3	25	9	75
Overall sample		69	1	1	68	99	12	17	56	81

Table 90 - Descriptive statistics for SJE DCS score bands by demographic groups

Protected characteristics		DCS bands								
		N	Low		Acceptable / good		Acceptable		Good	
		N	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*
Gender	*Males*	60	1	2	59	98	8	13	51	85
	Females	9	0	0	9	100	3	33	6	67
Ethnicity	*White*	61	1	2	59	98	8	13	51	85
	Other	8	0	0	8	100	3	38	5	63
Age	*up to 50 years*	57	1	2	56	98	9	16	47	83
	over 51 years	12	0	0	12	100	2	17	10	83
Overall sample		69	1	1	68	99	11	16	57	83

Table 91 - Descriptive statistics for SJE TLS score bands by demographic groups

Protected characteristics		TLS bands								
		N	Low		Acceptable / good		Acceptable		Good	
		N	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*
Gender	*Males*	60	1	2	59	98	8	13	51	85
	Females	9	0	0	9	100	3	33	6	67
Ethnicity	*White*	61	1	2	60	98	9	15	51	84
	Other	8	0	0	8	100	2	25	6	75
Age	*up to 50 years*	57	1	2	56	98	8	14	48	84
	over 51 years	12	0	0	12	100	3	25	9	75
Overall sample		69	1	1	68	99	11	16	57	83

The percentage of participants that scored each band did not differ significantly by gender, ethnicity or age. These results suggest that there is no issue with the fairness of the bands proposed.

7.4.2 Multi-Modal Interview

7.4.2.1 Scoring and banding

Scoring of the interview is based on the positive and negative evidence that arises during the interview.

Behavioural criteria

The interviewer is required to provide a score for each of the six topic areas using the indicator-based scale displayed in Table 92.

Table 92 - Rating scale for MMI topic areas

Rating	Meaning	
5	All of the positive indicators	No negative indicators
4	Majority of the positive indicators	No negative indicators
3	Less than half of the positive indicators	No negative indicators
2	None or some positive indicators	One negative indicator
1	None or some positive indicators	More than one negative indicator

Topic area scores are then averaged across each of the three main criteria, and bandings applied as shown in Table 93.

Note that due to the averaging process, it is possible to get a main criterion score that is not a whole number. The standard rounding conventions apply ie 3.5 would be rounded up to 4, and 3.49 would be rounded down to 3.

Table 93 - Banding scores on the MMI behavioural selection criteria

Main criterion average	Meaning	MMI Result
5	Exemplary: All positive indicators	PASS
4	Very good: Majority of the positive indicators	PASS
3	Acceptable: Less than half of the positive indicators	PASS
<3* (any topic area score <3)	One or more negative indicators	FAIL

* Score rounding process applies ie average score of 2.49 is rounded to 2 and 2.51 is rounded to 3

This means that no negative evidence is tolerated in the MMI for the behavioural criteria. The candidate fails the MMI if he or she provides negative evidence in the interview.

Communication

The interviewer is required to provide an overall score for communication based on the quantity and quality of positive evidence in line with the behavioural indicators. In the case of communication, each indicator is either rated '+' or '-'. The overall scoring key is shown in Table 94.

Table 94 - Rating scale for MMI communication

Rating	Meaning
5	All positive indicators demonstrated to an exemplary standard
4	All positive indicators demonstrated to a good standard
3	Most positive indicators are displayed to an acceptable level, training should improve the one or two minor instances of negative communication evidence
2	Some substantial examples of one or more negative indicators / difficult to understand in parts
1	Significant negative indicators / difficult to understand in the majority of the interview

A simple banding is then applied as outlined in Table 95.

Table 95 - Banding scores on the MMI communication criterion

Communication rating	Meaning	MMI Result
5	Exemplary: All positive indicators	PASS
4	Good: All positive indicators	PASS
3	Acceptable: minor points addressed through training	PASS
<3	Negative evidence / difficult to understand	FAIL

This means that only very minor negative evidence (eg use of one slang term – see the MMI manual for further examples) is tolerated in the MMI for the communication criterion.

The candidate fails the MMI if he or she shows substantial negative evidence which makes him or her difficult to understand.

7.4.2.2 Rationale and supporting evidence

This simple and transparent scoring approach provides a clear link between evidence provided in the interview, the score bands given, and the final pass/ fail decision. After consultation with the project steering group this it was decided that this was the most meaningful approach to scoring the MMI as it maintains a clear link to negative evidence.

Behavioural criteria

The pass / fail rule (any topic area score <3 = fail) means that a poor response on a single topic area is not concealed, and that negative evidence is not tolerated.

Communication

The pass/fail rule (communication score of less than 3 = fail) takes into account that minor communication issues (such as the use of slang or speaking too quickly) are trainable.

It was possible to estimate the impact of this scoring framework on the trial sample by substituting the topic area scores with the sub-criteria scores to produce main criteria scores. Across the sample of 91 participants there was one fail on Conscientiousness, one fail on DCS (both the same participant), and three fails on TLS.

Fairness

Table 96, Table 97, and Table 98 provide the frequency and percentage of members of each protected group category falling into each band and reaching an overall pass or fail.

Table 96 - Descriptive statistics for MMI Conscientiousness score bands by demographic groups

Protected characteristics		**MMI Conscientiousness bands**										
		N	**5**		**4**		**3**		**PASS**		**FAIL**	
		N	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*
Gender	*Males*	80	0	0	76	95	3	4	79	99	1	1
	Females	11	0	0	11	100	0	0	11	100	0	0
Ethnicity	*White*	81	0	0	77	95	3	4	80	99	1	1
	Other	10	0	0	10	100	0	0	10	100	0	0
Age	*up to 50 years*	80	0	0	77	96	3	4	80	100	0	0
	over 51 years	11	0	0	10	91	0	0	10	91	1	9
Overall sample		91	0	0	87	96	3	3	90	99	1	1

Table 97 - Descriptive statistics for MMI DCS score bands by demographic groups

Protected characteristics		MMI DCS bands										
		N	5		4		3		PASS		FAIL	
		N	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*
Gender	*Males*	80	0	0	11	96	2	3	79	99	1	1
	Females	11	0	0	11	100	0	0	11	100	0	0
Ethnicity	*White*	81	0	0	78	96	2	3	80	99	1	1
	Other	10	0	0	10	100	0	0	10	100	0	0
Age	*up to 50 years*	80	0	0	78	98	2	3	80	100	0	0
	over 51 years	11	0	0	10	100	0	0	10	91	1	9
Overall sample		91	0	0	88	97	2	2	90	99	1	1

Table 98 - Descriptive statistics for MMI TLS score bands by demographic groups

Protected characteristics		MMI DCS bands										
		N	5		4		3		PASS		FAIL	
		N	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*
Gender	*Males*	80	0	0	63	79	14	18	77	96	3	4
	Females	11	0	0	11	100	0	0	11	100	0	0
Ethnicity	*White*	81	0	0	66	82	13	16	79	97	2	3
	Other	10	0	0	8	80	1	10	9	90	1	10
Age	*up to 50 years*	80	0	0	66	83	11	14	77	96	3	4
	over 51 years	11	0	0	8	73	3	27	11	100	0	0
Overall sample		91	0	0	74	81	14	15	90	97	3	3

Three t-tests were conducted to explore whether there was a significant difference in protected characteristics of those who passed / failed each main criterion. No significant differences were found for any of the main criteria according to gender, age or ethnic group. It was not possible to perform meaningful statistical tests to examine the relation between MMI behavioural criteria score bands (5/4/3/fail) and key demographic characteristics due to the small number of cases in each band.

Verbal communication

The rating scale used to evaluate Verbal communication during the trial was changed in response to issues with the validity of the MMI as a measure of communication, and feedback from interviewers involved in the trial. Rather than relating simply to the proportion of indicators, the Verbal communication scale now refers to the quality of evidence demonstrated (please see Table 94 above).

The results in Table 99 are therefore just an indication of the fairness of the new recommended MMI scale, based on the findings from the trial.

Table 99 - Descriptive statistics for MMI Verbal communication score bands by fairness groups

Protected characteristics		**MMI Verbal communication bands**											
		N	5		4		3		PASS		FAIL	
			Freq.	*%*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*
Gender	*Males*	80	0	0	79	99	1	1	80	100	0	0
	Females	11	0	0	10	91	1	9	11	100	0	0
Ethnicity	*White*	81	0	0	79	98	2	2	81	100	0	0
	Other	10	0	0	10	100	0	0	10	100	0	0
Age	*up to 50 years*	81	0	0	80	99	1	1	81	100	0	0
	over 51 years	10	0	0	9	90	1	10	10	100	0	0
Overall sample		91	0	0	89	99	2	1	91	100	0	0

No participants in the current trial failed on the MMI Verbal communication scale. As before, it was not possible to perform meaningful statistical tests to examine the relation between MMI behavioural criteria score bands (5/4/3/fail) and key demographic characteristics due to the small number of cases in each band.

Overall, the fairness analysis suggests that there is not an issue with the fairness of the pass/fail banding of MMI scores. It has not been possible to conduct meaningful comparisons between the different score bands (5/4/3/fail) for the different protected characteristics due to the limited number of people in each group.

7.4.3 Combining the SJE and MMI to make a final pass/fail decision on behavioural criteria

7.4.3.1 Sample

A total of 41 participants sat both Version A of the SJE and the MMI. All of these participants had behavioural performance data, 36 had communication data, 29 had all elements of driving data, and 16 had training data.

7.4.3.2 Incremental validity

As shown in Table 100, the only SJE and MMI main criterion scales to be significantly correlated were the two Conscientiousness scales.

Table 100 – Correlations between SJE and MMI scores

MMI main criteria scores		**SJE Version A – Main Criteria Scores**		
		Conscientiousness	*DCS*	*TLS*
Conscientiousness	*r*	.30		
	p	.02		
	N	44		
DCS	*r*		.08	
	p		.31	
	N		44	
TLS	*r*			-.04
	p			.40
	N			44

Although both scales were found to have good criterion validity, linear regressions were carried out to establish whether the MMI Conscientiousness scores accounted for any variance in the criterion scores above and beyond that accounted for by the SJE.

The regressions showed that together, the SJE and MMI accounted for 23% of variance in the managers' ratings of conscientiousness. SJE Conscientiousness was associated with a significant increase in the manager rating of conscientiousness (p =.02), but the MMI Conscientiousness score became insignificant (p =.20) once the SJE was accounted for.

In other words, the analysis suggests that the MMI conscientiousness score might not be adding value. However, it was decided that it is appropriate to retain both the SJE and MMI as a) the other main scale scores are not correlated and b) keeping both measures will improve the construct validity as they use different methods to provide a measure of behavioural preference.

As discussed in the separate SJE and MMI sections, correlations between the SJE and MMI and all other assessment method scores produced no more significant correlations than would be expected by chance, suggesting good incremental validity.

How the scores are combined

For each of the three main behavioural criteria the possible score bands are:

- 'Moderate/good' or 'low' on the SJE
- '5', '4', '3' or 'Fail' on the MMI

The scoring matrix shown in Table 101 is used to combine the SJE and MMI scores for each of the main criteria:

Table 101 - Matrix for the combination of SJE and MMI scores

MAIN CRITERIA MATRIX		**MMI rating**			
		5	*4*	*3*	*Fail*
SJE rating	*Moderate / good*	P	P	P	F
	Low	P	P	F	F

This means that a Fail on the MMI equates to an overall fail of the behavioural selection criteria, but that a 'low' result on the SJE can be compensated for by a rating of 5 or 4 on the MMI.

7.4.3.3 Fairness

Chi-squared tests were performed to examine the relation between combined SJE/MMI score bands and the overall pass/fail decision for key demographic characteristics. Differences between the key demographic groups for the three combined SJE MMI scores are shown in Table 102, Table 103, and Table 104. Note that the results of these tests should be interpreted with caution given the small number of cases in each band (ideally there would be at least five cases in each cell of the test).

Table 102 - Overall Conscientiousness pass and fail rates based on the combined SJE MMI score

			Overall Conscientiousness Pass		Overall Conscientiousness Fail	
Protected characteristic		*N*	*Freq.*	*%*	*Freq.*	*%*
Gender	*Males*	*38*	37	97	1	3
	Females	*6*	6	100	0	0
Ethnicity	*White*	*39*	38	97	1	3
	Other	*5*	5	100	0	0
Age	*Age up to 50 years*	*36*	36	100	0	0
	Age over 51 years	*8*	7	88	1	13
Overall sample		44	43	98	1	2

Table 103 - Overall DCS (DCS) pass and fail rates based on the combined SJE MMI score

			Overall DCS Pass		Overall DCS Fail	
Protected characteristic		*N*	*Freq.*	*%*	*Freq.*	*%*
Gender	*Males*	38	37	97	1	3
	Females	6	6	100	0	0
Ethnicity	*White*	39	38	97	1	3
	Other	5	5	100	0	0
Age	*Age up to 50 years*	36	36	100	0	0
	Age over 51 years	8	7	88	1	2
Overall sample		44	43	98	1	2

Table 104 - Overall TLS (TLS) pass and fail rates based on the combined SJE MMI score

			Overall TLS Pass		Overall TLS Fail	
Protected characteristic		*N*	*Freq.*	*%*	*Freq.*	*%*
Gender	*Males*	38	35	92	3	8
	Females	6	6	100	0	0
Ethnicity	*White*	39	37	95	2	5
	Other	5	4	80	1	20
Age	*Age up to 50 years*	36	33	92	3	8
	Age over 51 years	8	8	100	0	0
Overall sample		44	41	93	3	7

The percentage of participants that passed or failed the three SJE/MMI main criteria did not differ significantly by gender or ethnicity. The results indicated that it is possible that candidates may be more likely to fail Conscientiousness or TLS if they are aged 51 or over, but with only eight people over 51 it is difficult to draw and firm conclusions.

The overall pass and fail rates, taking into account the three combined SJE MMI scores are shown in Table 105.

Table 105 - Overall pass and fail rates for the complete behavioural criteria

			Overall Pass on behavioural		Overall Fail on behavioural	
Protected characteristic		*N*	*Freq.*	*%*	*Freq.*	*%*
Gender	*Males*	38	34	90	4	11
	Females	6	6	100	0	0
Ethnicity	*White*	39	36	92	3	8
	Other	5	4	80	1	20
Age	*Age up to 50 years*	36	33	92	3	8
	Age over 51 years	8	7	88	1	13
Overall sample		44	40	91	4	9

The results suggest that there is no cause for concern about the fairness of the combined SJE MMI scores. However, due to the limited sample size it is particularly important that fairness is revaluated on an on-going basis as part of the future reviews.

7.4.4 Overall pass rate for the behavioural assessment

Overall, in this trial 91% of the participants were classified as passing the behavioural assessment process. When the SJE and MMI are implemented, a lower pass rate is expected because - as outlined earlier in this section – the participants in the sample were all trainees or drivers and had reasonable behavioural ratings from their managers. The SJE and MMI have not been assessed with a candidate population so it is not possible to predict what the pass rate will be.

7.5 Development of the assessment process for written communication

7.5.1 Recommended scoring and banding

Instead of being used as a pass/ fail measure, it is recommended that the WCT is used to provide a qualitative assessment of written communication. It can be used to identify training needs and specifically which sub-areas candidates may require particular attention to during communication training.

Five scores are produced on the WCT, a score for each of the four sections and an overall score. Each section score is banded into one of three bands: ‘low’, ‘moderate’ and ‘good’.

These also apply to the overall score for the WCT with an additional 'illegible' band. These are explained further in the tables below.

Table 106 - Explaining the banding on the WCT

WCT section	Scoring and banding	Notes	
Legibility	Scores between 0 and 2 are obtained. = 0: Low = 1: Moderate = 2: Good	Candidates must meet ≥ 1 (ie have mostly or completely legible handwriting) to obtain an overall WCT score of moderate or good.	If a candidate does not reach the 'Good' level on accuracy, written comprehension or structure, a qualitative comment is added to the scoring sheet clarifying the weaker area(s).
Accuracy	Scores between 0 and 6 are obtained. ≤ 1: Low = 2-3: Moderate ≥ 4: Good		
Written comprehension	Scores between 0 and 10 are obtained. ≤ 1: Low = 2-3: Moderate ≥ 4: Good		
Structure	Scores between 0 and 4 are obtained. (The score is comprised of up to 2 points that are awarded for logical sequence and up to 2 points that are awarded for relevance). = 0: Low = 1: Moderate ≥ 2: Good		

Overall score and bands
The maximum total number of points available is 22. The total WCT score is provided alongside the banding on the final scoring sheet for assessors. • **Good** = Candidate achieves good on all sections. This candidate's total score indicates the quality of written comprehension they have. It is unlikely to require special attention during communication training. • **Moderate** = Candidate achieves moderate scores on one or more of the WCT sections. This candidate may require some attention during communication training compared to other candidates. The areas of training are indicated by the areas highlighted as moderately scored. • **Low** = Candidate achieves a low score on one or more of the WCT sections. This candidate may require extensive training in certain aspects of written communication than others (as highlighted by the qualitative comments) compared to other candidates. • **Illegible** = Candidate scores low on legibility, regardless of the score on other sections. This is not likely other than in special circumstances so should be explored further in conversation with the candidate to understand any underlying issues.

The recommendation is for both versions of the WCT to be implemented. The WCT version should be randomly assigned with equal spread across candidates in any one sitting (ie aiming for half of the candidates taking version 1 and the other half version 2).

7.5.2 Rationale and supporting evidence

It was not considered appropriate to make the WCT a pass/fail measure because written communication is not a time critical safety related task for train drivers. The WCT was developed on suggestion from the industry steering group, to assess written communication in a more structured way and it did this well. However, it was not working as a proper psychometric test because it assessed too many different aspects of writing skill with too few items, despite good validation evidence.

The WCT expectancy tables could not be used to provide a clear indication of suitable score bands because of limited range in performance scores for the participants, and therefore regression analyses could not be used to produce plausible associated scores. For this reason, the score bands were set in relation to where drivers and trainees had low scores on the WCT. These scores were standardised (creating z100 scores).

The banding limits for each section and for the overall score have been set to represent that drivers involved in the trial had good performance ratings. The 'good' band has been set at the minimum levels reached by almost every candidate in the T948 trials. The 'moderate' band has been set at a level to reflect a score that is close to good but with some weakness. The 'low' band was set at two standard deviations below the mean to reflect the position of very poor performance.

The ‘illegible’ band of the overall score has been set using the rationale that in order for a candidate to have a basic level of written communication their writing must be legible to others.

The WCT is now scored using a scoring framework that has been improved on the basis of the results from the T948 trials. The scoring framework was amended to address the issues raised (see appendix B) and in particular the revised versions allow candidates to gain more marks from the accuracy, comprehension and concise structure sections of the test. Some amendments were also made to each storyboard and associated forms based on feedback and analysis from the trials.

7.5.3 A note on the interim score bands

The candidate sample is not expected to have the same distribution and so as more data are collected these assumptions can be tested, and the score bands adjusted accordingly. The development of a complete norm group will also allow the final results to be presented as percentiles.

With the collection of more data to produce a norm group, further changes could be made to the WCT and if there is strong statistical support for differentiating between the bands, this distinction could be used by companies to help make final scoring decision.

7.5.4 Fairness

It was not possible to rerun a full analysis on the final versions of the WCT due to the changes that were made to the scoring. However, the differences amongst the protected characteristic groups that were found in the original analysis were taken into account in the revised versions and it is expected that there will be no difference in WCT scores according to gender, ethnicity or age.

7.5.5 Incorporating the WCT into the selection process

The WCT scores are banded as good, moderate, low or illegible and indicate which candidates have better written communication skills than others. The WCT score does not have an explicit pass/fail minimum mark as it is not a time-critical safety related aspect of the train driver role and consequently, the scores are not combined with the MMI verbal communications score to provide an overall assessment of a candidates communications skills.

Instead, the WCT score should be treated as a ‘footnote’ to the selection process that may help assessors to identify training needs.

If a candidate has performed well in all other respects except written communication then it is recommended to have a conversation with them to try and understand the issues.

7.6 Estimating the pass-rates for each selection criterion

Where available, the norm tables for each assessment score in the recommended process were consulted to estimate how many

people in the candidate pool might fail each score. However, practice in the industry assessment centres is to sift out candidates during the assessment day if they fail parts of the process. Therefore, assessments that are later in the process are only taken by candidates who have passed preceding assessments. Regression analyses were used to estimate the amount of shared variance between each assessment score and the assessment scores of preceding assessment methods. The failure rate was adjusted to take into account this shared variance. This was used to predict a failure rate for each assessment method and each selection criterion. This was only possible for the cognitive and psychomotor assessment methods where norm data of some sort was available. No norm data was available for the SJE, MMI or WCT because they have not been trialled on a candidate group.

It is important to note that these estimates are very tentative because they are based on norm groups provided by the test publishers and not proper candidate norms. In addition, the sample size of the trial data was small so the results of the regression analysis could be misleading. The pass rate will need to be monitored if the recommended process is implemented to check that it is not at an unreasonable level.

Table 107 – Estimated pass rates for each cognitive and psychomotor assessment method

Assessment method	Percentage excluded of all those who sit it	Cumulative percentage excluded during assessment day	Additional percentage of total candidate pool excluded
TEA-Occ	18.4	17.6	17.6
Paper Group Bourdon	21.0	34.0	16.4
TRP1	5.5	37.6	3.6
TRP2	5.5	41.0	3.4
WAFV	16.5	52.5	11.5
TAVTMB	7.1	55.9	3.4
2HAND	4.7	57.9	2.0

Table 108 – Estimated pass rates for each selection criterion

Selection criterion	% excluded of all those who sit it	Cumulative % excluded during assessment day	Additional % of total candidate pool excluded
Attention	39.4	34.0	34.00
Memory	5.5	37.6	3.6
Reasoning	5.5	41.0	3.4
Vigilance	16.5	50.4	9.4
Reaction time	4.2	52.5	2.1
Perception	7.1	55.9	3.4
Hand-coordination	4.7	57.9	2.0

7.7 Overall fairness analysis based on trial sample

Every effort was made to collect information from females, older candidates and ethnic groups. However, people from these demographic groups are so poorly represented in the train driver population that it was only possible to obtain a very small sample. In addition, data from the TRP and Group Bourdon were collected from the RACF database from 2010/2011 which represented over a thousand people. Test manuals were also reviewed to look at group comparisons on age, gender and ethnicity.

Assessment method scores were compared between groups to check whether the recommended pass marks are likely to result in adverse impact for minorities. The benchmark for this judgement is whether the pass rate for the minority group exceeds 80% of the pass rate for the majority group.

However, the small sample sizes in the minority groups mean that a change of just one person results in a huge increase in the percentages failing. This is why we have had to group into white and non-white which is a limitation of the study.

The recommended process has been designed to be fair and to address the problems with adverse impact that are present in the current process. Further evaluation will be needed to confirm if this has been successful when the new process has been in use for some time. It should also be noted that good preparation materials are also key to ensuring the process is fair.

The pass/fail rates for the overall assessment process as applied to the trial sample are shown in Table 109. Not all participants completed all assessment methods so it was necessary to make an assumption that participants would have passed any assessment methods that they did not have scores for. Appendix I contains a full table of the pass rates of different protected characteristics groups for each individual assessment score when the recommended

scoring rules were applied to the trial data or previous assessment centre data where available. It is not possible to draw firm conclusions about fairness from this data because the numbers of people in the minority groups are so small. However, it does show that fairness has been considered at every possible stage. There is no reason to believe that there will be a problem with fairness if the recommended process is implemented but it still needs to be monitored to confirm this.

Table 109 – Pass/fail rate if recommended scoring rules are applied to the trial sample

Pass rates for trial participants from different groups (%)							
Group		*Pass %*	*Pass n*	*Fail %*	*Fail n*	*Passes 4/5ths rule*	*Number of assessments (n total)*
Ethnicity	*White*	88	310	12	44	Yes	354
	Other ethnicity	70	16	30	7		23
Gender	*Males*	90	324	10	37	Yes	361
	Female	88	21	12	3		24
Age	*50 and under*	91	303	9	30	Yes	333
	51 and above	80	41	20	10		51

8 Appendices

A. Selection of cognitive and psychomotor assessment methods

A.1 Introduction of vigilance and hand coordination as additional selection criteria

It was recommended in project T340 (RSSB, 2005) that the selection criteria by which train drivers should be assessed needed to be reviewed in light of the legislative changes including the Driver Licensing Directive and anticipated changes to the driver role. This review identified that the selection criteria assessed in Great Britain needed to be updated which was agreed with industry representatives. These criteria included some behavioural criteria which are discussed in other sections and also some important cognitive criteria including vigilance and hand coordination. This appendix explains which tests were shortlisted to be trialled for the assessment of vigilance (section A.2) and hand coordination (section A.3).

A.2 Evaluation of vigilance tests

There is a misconception that is common, which is that sustained attention and vigilance are terms that can be used interchangeably. However, they are not the same. Vigilance can be defined as the ability to attend and respond to stimuli which occur relatively infrequently and over extended periods of time (Robertson, 2004). Sustained attention, on the other hand, can be defined as the ability to attend to stimuli that are presented fairly frequently. This is an important distinction to make in the assessment of vigilance in the train driver selection process (RSSB, 2005). In addition, studies have confirmed that a lack of vigilance is a contributory factor in train driving incidents and accidents (Robertson & Garavan, 2004).

This definition of vigilance was considered when identifying and shortlisting potential vigilance tests for the selection of train drivers. The TEA-Occ lottery assessment method was initially considered for the assessment of vigilance but the T628 trials reported that it was not considered to be suitable for the assessment of vigilance. Therefore it became necessary to identify and shortlist alternative vigilance tests so that they could be evaluated as part of the T948 trials.

Ten tests were identified which were available as ‘off-the-shelf’ tests. These tests were evaluated based on the task the candidate would be required to undertake and its relevance to the rail industry, the variables used for scoring the method, the administration time, available norm groups, use in other countries and evidence available on the psychometric properties of the methods, particularly construct validity, criterion validity and reliability. The summarised review of each test is illustrated in Table 110.

Table 110 - Evaluation of vigilance tests available

Test	Publisher	What it involves doing	Scoring	Test forms available	Administration time	Norm groups	Use in other countries	Reliability, validity	Advantages	Disadvantages
Continuous attention (DAUF)	Schuhfried	Triangles appear in a row onscreen, when a predefined number of triangles point down the candidate must press a button. How regularly the predefined number of triangles pointing down appear is random	Sum correct, mean time correct, sum incorrect, mean time incorrect	3 test forms which differ in number of triangles and how often the rows change. Only test form 3 is appropriate (1&2 designed for those with impaired attention). S3 has irregular intervals and presents 7 triangles at a time	35 minutes for S3 (includes a practice phase; the candidate receives no feedback as to the correctness of their answers in the practice)	S3 general population norm only, N=568	Austria, Germany	Cronbach's alpha between 0.64 - 0.99 depending on test form. Convergent validity (Cognitrone) = 0.53, construct validity assumed	Exactly the same keyboard cover can be used for this test as with the 2Hand and other VTS tests trialled	Test does not require the inhibition of behaviour so may not be suitable for the assessment of vigilance
SCAAT	OPC	The test uses a cancellation task like that in the Group Bourdon. It has three sections each covering one of the aspects of attention. Section 1 involves candidates searching for a single target shape, section 2 searching for two target shapes at any one time and section 3 searching for two target shapes at any one time but one shape is constantly changing.	Ability to maintain concentration on monotonous tasks, multi-tasking, attentional switching	One	This is a paper and pencil test. The time limit on the test is 21 minutes but total administration time is about 45 minutes.	13 norm groups available including UK and Australian train drivers	Currently used in UK by some TOCs	Validated against train driver training	This test seems often to be used as a companion test to the Group Bourdon and gives an alternative assessment of attention. These aspects are assessed in three different parts of the test.	It aims to measure three of the four main aspects of attention, namely concentration (or focussed attention), divided attention and attention switching. But may not be suitable for the assessment of vigilance

Test	Publisher	What it involves doing	Scoring	Test forms available	Administration time	Norm groups	Use in other countries	Reliability, validity	Advantages	Disadvantages
MICROPAT battery	UK Armed Forces	Landing subtest - candidate has to land a plane on a runway	Landing subtest - measures attention	Not known	Not known	Army pilots?	Used by a number of commercial airlines and military organisations	Criterion validity (pilot training)	Based on a job/work sample approach.	May not be for commercial use, does not necessarily measure vigilance. Would require standardisation
VIGIL	Schuhfried	Brightly flashing dot travels along a circular path in small jumps. Occasionally the dot does a double jump, and the candidate should press a button in response to this	Main scoring variables - number of correct reactions, number of incorrect reactions, mean value of reaction time for correct reactions	3 test forms	Approximately 30 minutes	Car drivers N=143 (test form 1), Swedish applicants for technical occupations N=367, Portuguese pilots N=178 (both test form 2)	Not known	Reliability is 0.82. Test score significantly correlate with attention performance in brain injured patients.	Exactly the same keyboard cover can be used for this test as with the 2Hand and other VTS tests trialled	
WAFV	Schuhfried	Candidates are required to respond to changes in a square that is presented on screen and which changes colour at irregular intervals	Main scoring variables – number of missed reactions, mean reaction time, number of false alarms	Four test forms	Approximately 30 minutes		Not known	Reliability is 0.96. Construct validity – WAFV significantly correlates with other tests of attention. Loads onto vigilance factor in principal components analysis of other tests of attention.	Exactly the same keyboard cover can be used for this test as with the 2Hand and other VTS tests trialled	

Test	Publisher	What it involves doing	Scoring	Test forms available	Administration time	Norm groups	Use in other countries	Reliability, validity	Advantages	Disadvantages
CompACT	Hogrefe	Candidates must continuously monitor a radar screen and identify when there are more circles than squares visible.	Response times across a fixed period are compared to identify changes in concentration levels	More complex version of the Concentration test	Up to 25 minutes		Perhaps in GB by Network Rail		Currently trialled by Network Rail Either touch screen or run from a computer programme/CD Very accurate time measurement (milliseconds) Reasonable face validity	Test manual not available in English (only in German). Limited technical support in the UK.
Vigilance test	Vilis	Not known	Observation acuteness, sense of detail when under acoustic and visual stress, measures reaction time, quality, variation when searching for an item within a group of items	One	Not known	Not known	Not known	None provided		Website and standard of English extremely poor
Psychomotor or Vigilance Task	NASA	Designed for astronauts to complete at various times eg when they are about to carry out a difficult task. Involves pressing buttons in response to visual stimuli on a hand held device	Changes in psychomotor speed, lapses in attention	One	3 minutes	Astronauts	None	No statistical information available	Easy to use, portable device, short administration time	Cannot be computerised, new equipment would have to be purchased, may not be for commercial use, no appropriate norm groups

Test	Publisher	What it involves doing	Scoring	Test forms available	Administration time	Norm groups	Use in other countries	Reliability, validity	Advantages	Disadvantages
Test of attentional vigilance (TOAV)	PEBL	2 test conditions with geometric stimuli appearing either at the top or bottom of the screen; half the test is infrequent (the target/non-target ration is 1:3.5) thus candidate must pay close attention, and it's fatiguing/boring. 2nd half is frequent (3.5:1) so candidate has to respond most of the time and thus inhibit response occasionally	Reaction time (milliseconds), attention, response time variability, errors of commission, errors of omission, d score, ADHD score	One	22 minutes	General population	Not known	None provided		The free version of the TOVA, thus too risky to use as no technical support
Test of variables of attention (TOVA)	The TOVA Company	2 test conditions with geometric stimuli appearing either at the top or bottom of the screen; half the test is infrequent (the target/non target ration is 1:3.5) thus candidate must pay close attention, and it's fatiguing/boring. 2nd half is frequent (3.5:1) so candidate has to respond most of the time and thus inhibit response occasionally	Reaction time (milliseconds), attention, response time variability, errors of commission, errors of omission, d score, ADHD score	2 (visual and auditory versions)	22 minutes	General population		Discrimination ability 0.80 sensitivity and 0.80 specificity	Free technical support, free software updates	Clinical use only thus standardisation needed, technical support only available during 9am - 5pm US time

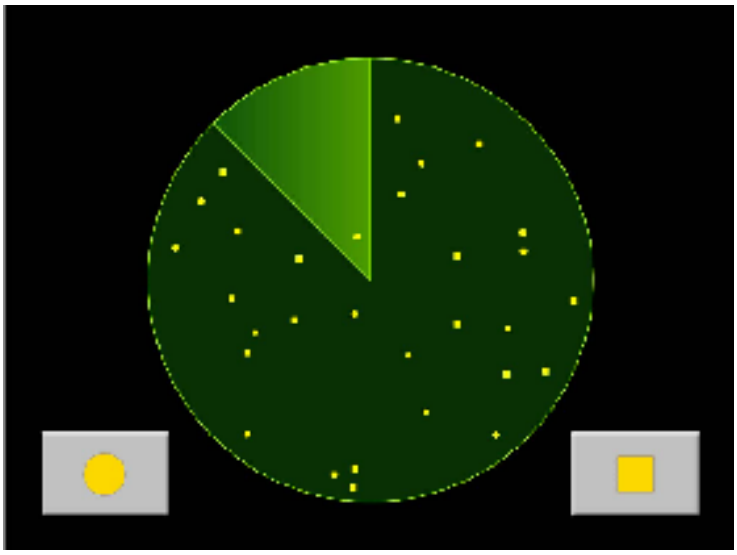
The evaluation of these methods considered that whilst potentially useful measures of attention, there was not sufficient evidence to support the use of some methods as providing a measure of vigilance specifically. These methods included SCAAT, MICROPAT, DAUF (see Figure 4) and NASA's Psychomotor Vigilance Task and were not considered further.

Figure 4 - DAUF

Other tests were not available in English such as Vilis, which were also not considered further.

Other assessment methods were more appropriate for clinical patients, including TOVA and TOAV, which were not considered further as standardisation would be required to render these tests suitable for occupational use. These methods were also considered to have poor face validity. In addition, other assessment methods were available with only limited technical support which has been reported as a problem with current tests such as the DTG. CompACT (the Computerised Attention and Concentration test) (Figure 5) was considered further as the target stimuli are presented infrequently and randomly which adheres to Robertson's theory of vigilance and Network Rail also trialled CompACT for the assessment of vigilance in signallers.

Figure 5 – CompACT



However, at the time of this review, Hogrefe were only able to provide limited technical support for CompACT. Due to dissatisfaction with the lack of technical support with some methods in the current assessment process, the T628 Steering Group did not wish to further pursue CompACT.

Two vigilance assessment methods – WAFV and VIGIL – were shortlisted for trial based on their adherence with the accepted definition of vigilance, that is, the candidate must respond to stimuli which are presented infrequently, the available norm groups and the psychometric properties of the methods.

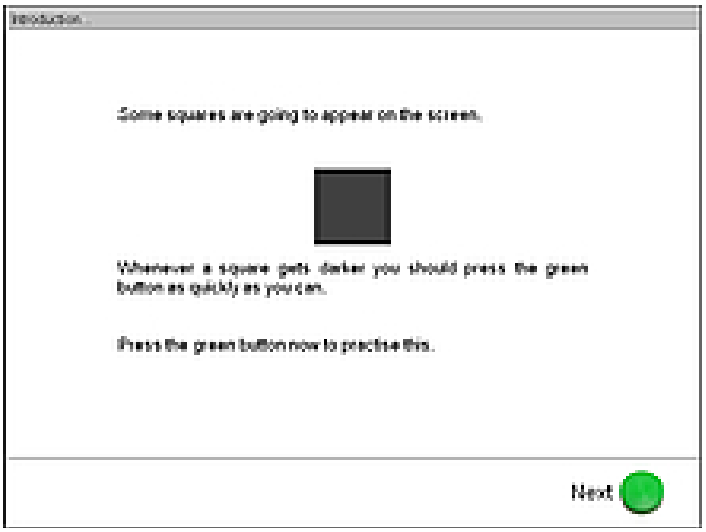
Shortlisted vigilance tests

WAFV and VIGIL are described further and the test forms are also described.

WAFV

The WAF test battery consists of six tests which assess various sub-functions of attention (Sturm, 2006). The WAFV is one of these tests and can be administered independently to assess vigilance. Like VIGIL, it is based on studies by Mackworth where stimuli are presented at very irregular intervals and at a very low frequency compared to the number of irrelevant stimuli (Sturm, 2006). This can also be compared to the train driving task, where drivers must respond to cautionary signals or other hazards which occur infrequently.

Figure 6 – WAFV



There are four test forms available within the WAFV test. These forms differ in the construction and stimulus conditions. The candidate is either presented with visual or auditory stimuli, and those stimuli also differ in frequency. The candidate must respond to these stimuli which occasionally diminish in intensity (either becoming darker or quieter). When sustained attention is being measured, the stimuli make up 30% of the stimuli, whereas when vigilance is being measured, they constitute 5% of the stimuli.

Table 111 - Construction and stimulus conditions of the WAFV test forms

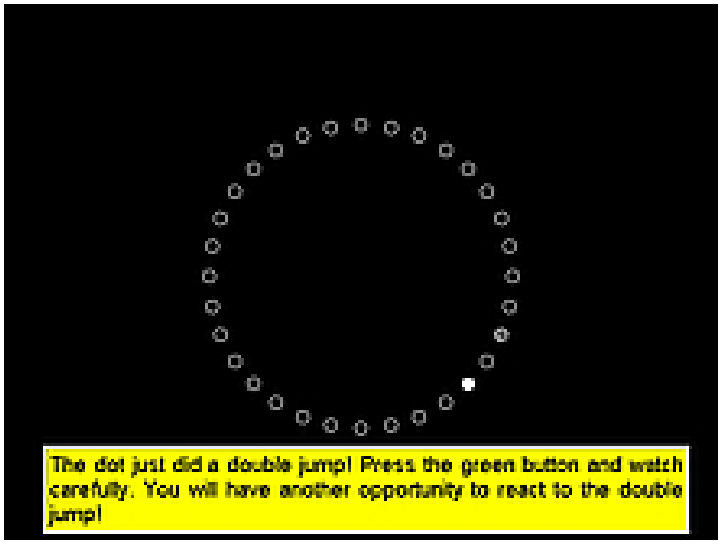
Test form	Type of attention	Type of stimuli	Relevant signals	Frequency
S2	Vigilance	Visual	Square becomes darker	5%
S4	Vigilance	Auditory	Sound becomes quieter	5%
S6	Sustained attention	Visual	Square becomes darker	30%
S8	Sustained attention	Auditory	Sound becomes quieter	30%

For train driver assessment, test form S2 was selected for trial as the stakeholder group considered the presentation of visual stimuli (as opposed to auditory stimuli) more relevant to the train driving role, particularly regarding responding to visual warning signals. The frequency of stimuli which is best suited to the driving role is 5% as warning signals may occur fairly infrequently on the mainline railway.

VIGIL

VIGIL (Puhr, 2004) is based on the theory that suggests target stimuli should be presented with low frequency and randomly (irregular stimuli being harder to predict and therefore attend to than regularly presented stimuli). This encourages sustained watchfulness in a stimulus deficient situation which adheres to the accepted definition of vigilance. In this test, a brightly flashing dot travels along a circular path in small jumps. Occasionally, the dot takes a double jump which the candidate must respond to by pressing a button on the control panel. In the chosen test form, the outline of the circular path is not shown; instead, the candidate must monitor the movement of the dots only.

Figure 7 - VIGIL



There are three test forms which differ in the type of visual stimuli presented:

Table 112 - VIGIL test forms

Test form	Type of stimuli	Application	Time
S1 Quatember-Maly	The individual dots of the circular path are shown on the monitor as small circles. This form is suitable for people whose attentional capabilities are significantly below average.	People with severe attentional deficiencies.	30 mins
S2 Müggenburg 33	In this form, no circular path is shown in the monitor. The respondents have to estimate whether the flashing dot has taken a double jump or not. The number of stimuli presented is less than S1.	People with normal attentional abilities.	35 mins
S4 Müggenburg 66	The same as test form S2 but the testing time is increased to 70 mins.	People with normal attentional abilities.	70 mins

Test form S2 Müggenburg 33 was selected for trial as the test form is suitable for people with normal attentional capabilities, whereas test form S1 is more relevant to people with attentional deficiencies and the administration time of test form S4 was considered too long to fit into the assessment day.

A.3 Evaluation of hand coordination tests

During the course of the T628 project, hand coordination was added to the selection criteria in response to the requirements of the European Commission Directive 2007/59/EC, now transposed into national law as TDLCR, 2010. A measure of hand coordination had not been included in the T628 trial so it was recommended to identify and evaluate a test for this purpose in future work.

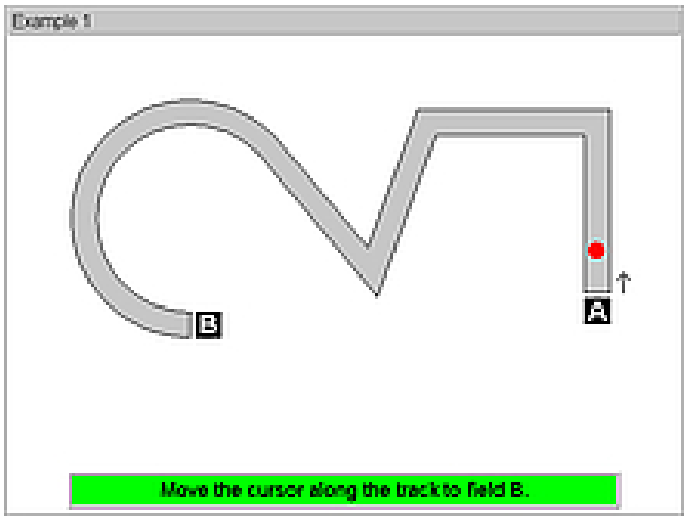
There was a growing body of evidence that the DTG, which is designed to measure the ability to operate hand and foot controls, lacked criterion validity in the assessment process, so alternatives needed to be identified. In order to identify suitable tests of hand

coordination, RSSB collected information from other European Member States in June 2009 on the selection criteria and assessment methods used.

A questionnaire was sent to European Member States who were represented on the Community of European Railways Psychologist's Subgroup. It was found that the 2HAND test from the Vienna Test System was used by ÖBB in Austria and CFL in Luxembourg, and later known that Queensland Rail and Queensland National in Australia also used this method. SNCF reported using an alternative version of the 2HAND test where handles need to be turned on the side of a box in order to move the ball round the track. It was therefore considered that the 2HAND test from the Vienna Test System should be trialled as other European Member States reported positive feedback for the test and had demonstrated that this method would adhere to the requirements of the European Commission Directive 2007/59/EC.

The candidate is required to make a red dot move along a given track using two joysticks. The track consists of three sections that make increasingly difficult demands on the coordination of the left and right hand. The point is moved from right to left.

Figure 8 – 2HAND



Five different test forms are available:

Table 113 - 2HAND test forms

Test form	Type of administration	Runs
S1	Joysticks	10 runs
S2	Joysticks	4 runs
S3	Control knobs	4 runs
S4	Control knobs	10 runs
S5	Joysticks (inverse direction assignment)	10 runs
S6	Joysticks (one-sided)	10 runs

Test form S1 was selected for trial as joysticks were considered as more suitable than control knobs for assessing the type of coordination required of train drivers.

A.4 Conclusion

A.4.1 Vigilance

Two vigilance tests were shortlisted for trial in the T948 study. These were WAFV (test form S2) and VIGIL (test form S2 Müggenburg 33).

A.4.2 Hand coordination

The 2HAND (test form S1) was shortlisted for trial to assess hand coordination.

Further information about WAFV, VIGIL and 2HAND is provided in Annex 3 Section 2 and the validation results of these assessment methods is provided in Annex 3 Section 5 of this report.

B. Process of development of the written communications test

B.1 Aims of the method

The WCT was designed to measure the ability to communicate effectively in writing. The test gathers evidence on four areas of written communication: accuracy, comprehension, legibility and structure. These areas were identified by SMEs as being important to the train driver role. The WCT was designed to be used together with the MMI verbal communication rating scale as a measure of the communication selection criteria.

B.2 Development process and rationale

B.2.1 Why use this type of method to measure the behavioural criteria?

Following on from recommendations in project T340 (RSSB, 2005), the WCT was developed in response to concerns raised by the industry steering group, during project T628, that the CBI candidate interview form did not offer a structured method of assessing written

communication skills. The current approach was not standardised and candidate responses on this form were variable. Therefore, the WCT was developed as a simple and quick assessment of written communication skills to take into account the types of reports that train drivers need to produce as part of their work.

B.2.2 The use of good practice and development decisions

The process of developing the test began with considering the various aspects of communication to be measured using the MMI and WCT. Although theories of communication were considered, it was felt that it would be more appropriate to consider SMEs' views about the relevant communication skills for the train driving job. This created a list of positive communication indicators which included a few aspects of written communication (see Table 115). One TOC also provided a list of communication indicators which informed the development of the WCT scoring key.

Table 114 displays the positive communication indicators that were developed with the SME group.

Table 114 – Positive communication indicators used in the development of the WCT

Positive communication indicators
Clear message
Concise
Logical structure
Comprehensible
Relevant answers to questions
Maintain performance
Written:
Legible
Addresses the question
completeness
Spoken:
Good pace
Clear pronunciation
Audible volume
Avoids slang
Active listening eg responds to points in question

Table 115 - Communication indicators from one TOC

Positive	Negative
Clear and concise in communication	Spoken English not easy to follow
Listens to questions and responds appropriately	Longwinded and verbose
Style of communication has good structure	Presentation structure unclear, answers do not really fit questions
Understands the benefits of communication and has an open and involving style	Communication style is static and monotone

Example written reports were discussed with SMEs for understanding and background information about what skills drivers need to be able to complete the forms. The group agreed these to be the four areas of written communication assessed by the WCT: accuracy, ability to summarise (written comprehension), legibility, logical sequencing and conciseness. Spelling and grammar were not considered to be important for the forms that drivers fill in. The WCT accuracy section is the only section in which spelling is assessed because the accuracy of factual information is a core component of the driver role.

The group agreed that the content of the test should be about an incident in a work-based transport context because this is similar to what drivers would report on. These suggestions were developed into two versions of the tests that are now recommended for use. It was agreed that each test would consist of story split into a six picture storyboard. The scenarios comprise of a schedule that includes key details, and represent a story with a beginning, an incident or hazard that disrupts the work and a consequence from that event. It was also agreed that the WCT stories should not be too similar to any of the scenarios in the SJE nor be unfair to any group with protected characteristics.

With regards to design and use, it was agreed that the test-taker should be able to clearly identify the details and the test be printed in a minimum font size 12 to account for those with difficulties reading small text. As the written communication skills required for the train driving role are currently not time-critical, it was agreed and recommended that 25% extra time would be provided for those candidates with dyslexia, which is in line with good practice (Hagan, 2009; Moody, 2009; London Fire Brigade, 2008). If in future, it is proven that some written communication tasks are time critical, this recommendation could be reviewed over time and the extra time could be removed. It was agreed that after the candidate had looked at the picture storyboard, it would not be removed away from them to ensure memory recall did not influence the ability write effectively.

B.2.3 Development of scoring

Central to the development of the WCT was a standardised and structured scoring key. This was developed to assess all five aspects of written communication for both versions of the test. These were split into details (accuracy), summary (written comprehension), legibility and structure (which comprised of logical sequencing and conciseness). If scores on each section were simply combined to make an evaluation of written communication performance, it would be possible for a candidate to score very low on some items but high on others, and still have a reasonable mark on the assessment. This would fail to account for some criteria that are more important than others; for example, legibility needs to meet minimum standard or it is irrelevant how well the candidate performs on the other aspects. Consequently, in the original version trialled, each section required a score that had to be reached to receive an overall 'acceptable' mark. A total score was also produced and scores were banded according to this total.

Together with corresponding instruction forms and scoring sheets, two versions of the WCT of roughly equivalent complexity and detail were created by a designer. These were trialled internally with 15 participants to decide an appropriate time limit and also amongst the SMEs and their colleagues to gain feedback. Following these trials, a few modifications were made which led to the versions that were trialled in this research.

B.2.4 Key iterations and refinements

B.2.4.1Addressing low score variance

Following the trials, the initial analysis showed that the WCT had good levels of validity. However, it also suggested that both versions of the WCT did not reach the required levels of internal consistency (Cronbach's alpha) for reliability. This led to further analysis of the difficulty levels of each item. Item facility calculations suggested that the majority of the test was too easy for the candidates; however the test did appear to differentiate between high and low performers.

To understand the structure of the test, a factor analysis was run and the results indicated that there were too many factors for the number of items in the test. Most importantly, the analysis showed a lack of score variance across all 16 items. Too many factors indicated that there is a lot of unique variance in the correlation matrix and this meant that the average correlations were likely to be low, and consequently, the coefficient alpha. This was confirmed by looking at the inter-correlations and was the reason why Cronbach's alpha was low on both versions.

The analysis showed that the low score variance was due to a variety of reasons. Several of the scoring items in the WCT had very high means. For example, two of the items in the details (accuracy) section items in version 2 had 100% correct responses. Such items have very low, or zero, variance and are, therefore, very unlikely to

(1) correlate well with other items, (2) contribute to coefficient alpha or (3) carry any weight in the scale score. These items were too easy for candidates, as confirmed by the item facility analysis. Half of the items in the written comprehension section provided all of the differentiation for the whole test. Also, because the test was given to qualified train drivers the scores were grouped towards the higher end of the scale (high pass rates) and therefore group homogeneity played a role. Another cause was that most of the scale variance could be related to administration instructions; fewer prompts and more detailed information about the story were required.

B.2.4.2Revised scoring framework

Conceptually, each section of the WCT is different from each other (see Annex 3 Section 5.5.3); this was also supported by the analysis and accounted for the low internal consistency (Cronbach's alpha) levels on both versions. This therefore gives weight to the idea of retaining the original method of splitting the scoring into separate sections that contribute to the overall score.

In order to address the reliability issues, and in particular the lack of score variance, an alternative score framework was designed for both versions to see if this could improve reliability. In the new framework candidates could score up to two points per item on all sections except the details (accuracy) section. This section was still dichotomously scored as the factual details can either be accurately presented or be error prone.

In response to the lack of score variance and difficulty in the details (accuracy) section all items in the accuracy section were made more complex and therefore require a more diligent transference from the storyboard to the report form by the candidate.

Due to the unique variance that the summary section contributed to the overall test, each item of the test was analysed further and this led to several recommendations for change:

Both versions

Item 1: was a removal (v1)/was a taxi pick-up (v2) was removed as it did not contribute to the overall score. It could also be argued that those candidates who did not provide the information may have assumed that as this was their job, they did not need to explicitly state what they are doing. *Item 2: scheduled collection time/ scheduled pick-up time)* and *Item 3: scheduled delivery time/ latest drop-off time* were considered to be accuracy rather than comprehension items as candidates had to replicate the information exactly to gain points. Therefore, these items were removed from the scoring and moved to the details section.

Version 1

Version 1 concerned a delivery lorry colliding with a box of broken glass. A common misinterpretation was that the box was dropped from the van and comments from candidates indicated that this picture was ambiguous. There was a knock-on effect from this picture for the interpretation of the consequences depicted in the remaining pictures. This was evident in *Item 7: damage to the tyre* where only 39 of the 69 candidates correctly identified this item and *Item 8: lorry can't be moved* where 15 candidates did not gain a point. Additional analysis was undertaken to look at how many candidates failed on both items 7 and 8. Only 39% (27/69) of candidates correctly identified both items, compared to 63% (46/73) for version 2 on the same items. With several negative correlations as well, the implication was that the misinterpretations were possibly a result of not understanding the item 6 picture and the consequences. There were also concerns about whether it was clear to notice that damage had actually occurred to the tyre due to the printing quality of the picture and therefore both pictures have been amended.

Version 2

No misinterpretations were noted from the candidates' perspective during the trials but the scoring key was amended where duplication of similar information (ie *Item 7: traffic jam* and *Item 8 stuck in traffic*) was present. In line with version 1, additional changes were also made so that more points could be gained from the summary (comprehension) section.

Minor changes to the scoring framework for the legibility and structure sections were made to reflect a clearer understanding of how each point should be awarded. The 'concise' structure section of scoring was changed and relabelled to 'relevant' structure because Information needs to be presented in an objective and straightforward manner with no fabrication. However, the aim was for information to be presented concisely as well so the test instructions will require the report to be written within a specified word limit. As driver report forms do not take this approach, it would be possible for key details to be potentially missed out if instructions to write concisely within the short timeframe were adhered to.

B.2.5 Evaluating the revised scoring framework

Where possible, the changes to the WCT were evaluated by rescoring a sample of the tests and reanalysing reliability.

It was not possible to re-score the details section using the new framework or to rescore version 1 because of changes to the pictures so these changes could not be evaluated.

However, the revised scoring framework could be applied to version 2 so a sample of these were re-scored and analysed. This analysis showed that the new scoring framework for the summary section did

not significantly improve internal consistency of the WCT, gave more weight to some points than was justifiable and was more complex for assessors. However, the revised relevant structure section worked well.

B.2.6 Conclusion

The evaluation of the new scoring framework led to the conclusion that there was no benefit to changing the scoring framework for the summary section that could justify the additional complexity. Therefore, the original scoring framework will be used with some modifications so that the maximum score for this section will be 11 with one point available per item. Changes to the pictures, details section and scoring of the structure section will all be progressed. Instead of a maximum total of 18 points, 23 points will be available.

The entire battery of changes to the WCT is expected to create larger score variance when used during the assessment process as it will be harder than the original versions. The candidate pool is also expected to have a greater range of performance than the driver population. If implemented, further evidence can be collected that could be used to improve the WCT even more.

C. Process of development of behavioural assessment methods

C.1 Situational Judgement Exercise

C.1.1 Aims of the method

As explained in Annex three section one, the Situational Judgement Exercise (SJE) has been designed to measure the behavioural criteria that have been identified by subject matter experts as being important to the train driver role.

The SJE was designed to be used with the new interview (Multi-modal interview or MMI) as another measure of the behavioural selection criteria.

C.1.2 Development process and rationale

Why use this type of method to measure the behavioural criteria?

As documented in earlier project reports, a range of options were considered when deciding what type of measure should be used with the interview. It was decided that an SJE would be appropriate for a number of reasons:

- SJEs can be used to measure job-related behavioural tendencies and some of the relevant personality constructs such as conscientiousness, agreeableness and emotional stability (eg McDaniel, Hartman, Whetzel, & Grubb, 2007; Motowidlo, Hooper and Jackson; 2006).
- SJEs have been shown to have good criterion validity (McDaniel, Morgeson, Finnegan, Campion, and Braverman, 2001; Lievens, Peeters, & Schollaert, 2006).
- Applicant reactions to SJEs are positive (Lievens, Peeters, & Schollaert, 2006).
- A preference was expressed by assessment centre staff and union members within the industry for a situation-based assessment of behaviour rather than a personality inventory employing self-assessment.

No suitable ‘off the shelf’ SJEs were available at the time (mid 2007) and so the decision was made to develop a bespoke SJE that would be designed specifically around the train driver selection criteria.

The use of good practice in development

To ensure that development of the SJE was rooted in best practice, various sources of guidance were used in developing the SJE including an SJE Development training course run by the Division of Occupational Psychology and Weekley and Polyhart's 2006 'Situational Judgement Tests: theory, Measurement and Application' book.

Scenarios and corresponding items were originally based upon information collected on critical incidents in the rail industry. However, it was later decided that this could favour applicants with existing rail experience. The situations were rewritten so that they related to the same behaviours but were set in an everyday context.

In line with Motowildo, Hooper and Jackson's (2006) model of situational judgement, the actions were designed to present a dilemma where the attractiveness of the action would be influenced by the individual's dominant personality traits or behavioural preferences. Taking this approach means that there is not always a clear 'correct' answer, as selecting a certain response may contribute to a good score on one criterion but not on another.

The theory and scoring underpinning the SJE was sense checked throughout development by consultation with a special working group (consisting of driver managers and recruitment staff), and through scoring benchmarking trials.

The test developers asked members of the working group to review the SJE scenarios and the proposed links to behavioural criteria, and to provide suggestions for further scenarios for inclusion.

The SJE was designed to be used with the MMI to make a pass/fail decision for each of the behavioural selection criteria. As all selection methods have relative strengths and weaknesses (eg there is evidence to suggest that SJEs can have good reliability and validity, and good incremental validity over cognitive tests, but can be prone to participant faking - Lievens, Peeters & Schollaert, 2007), this is the best way to ensure that the process is valid, reliable and fair.

Development of scoring

Once the hypothetical scenarios and actions were developed, weightings of between -3 and +3 were applied to each selection sub-criterion for every action. This weighting was based on the development team's judgement of what which sub-criteria each action was measuring.

For example, if an action was thought to be a strong positive indicator of dependability, a fair negative indicator of calmness under pressure, and to be unrelated to other criteria, it would have a loading of 3 on dependability, a loading of -1 on calmness under pressure and no other weightings.

The scale responses and loadings are then combined to produce a sum product score for each sub and main criterion. The consistency of responses within each criterion is also calculated, and then banded using information derived from Dunlap, Burke and Smith-Crowe's (2003) paper on inter-rater agreement indexes.

Refinement of scoring and the SJE

A number of drivers (n = 24), trainees (n = 31), and driver managers (n = 7) were selected for inclusion in the scoring benchmark review. These participants were selected on the basis of performance data provided by managers, with the aim of getting a good spread of performers.

The participants were asked to complete the SJE and a bespoke version of the NEO-PI-R (a version limited to the scales which are comparable to the driver selection criteria). The NEO-PI-R is a well-established questionnaire measure of the five major domains of personality, and it was possible to use items from this to simulate scales aligning with the selection criteria. Scores on the bespoke NEO-PI-R, and performance ratings provided by managers, were used as a benchmark to indicate which participants were 'good' and 'poor' for each of the sub-criteria.

The analysis of this data led to a refinement of the SJE scoring scale whereby the 'best' scale point for each action (as defined by responses given by participants with the best NEO-PI-R results and job performance) was identified. Using this information, the SJE programme converts the raw scores by taking into account its proximity to the 'best' scale point.

This scoring benchmark review also helped to identify which items were not working particularly well, and so these were removed from the two versions of the SJE that were trialled (Version A and version B).

C.1.3 Key iterations and refinements

Using only one version of the SJE

Version A and B of the SJE were developed to be equivalent (ie with the same number of items and similar weight loading patterns), but the trial analyses suggested that the two versions were not sufficiently similar. Although good in parts, the overall criterion validity and reliability of version B was not as strong as for Version

A. The results of factor analysis also suggest that the two forms had slightly different factor structures. For these reasons it was decided that only Version A of the SJE would be recommended for implementation.

The particularly effective parts of version B ('checking' and 'attitude to work and people' subscales) were integrated into version A and the addition of these items meant that some of the less effective items from version A could be removed.

Amendments to the 'checking' and 'attitude to work and people' sub-scales

Although the quality of Version A of the SJE was generally very good, the validity of the 'checking' subscale, and reliability of the 'attitude to work and people scale' did not reach desirable levels. However, these subscales on version B of the SJE had good reliability and validity and so some of these items were transferred across to version A. It is expected that this will improve the properties of the final version of the SJE.

The use of one rating scale rather than two

The existing literature on SJE development discusses the various pros and cons of the two main types of responses: knowledge-based responses, eg good / poor or helpful / unhelpful (McDaniel and Nguyen, 2001; McDaniel et al, 2007) and behavioural tendency responses, eg likely / unlikely (McDaniel, Whetzel, Hartman and & Nguyen, 2006; McDaniel et al, 2007). Figure 9 provides examples of each scale.

How helpful is this response?

Very helpful	Helpful	Neither helpful nor unhelpful	Unhelpful	Very unhelpful
1	2	3	4	5

How likely or unlikely would you be to take this action?

Very likely	Likely	Unlikely	Very unlikely
1	2	3	4

Figure 9 - Examples of helpful / unhelpful and likely / unlikely scales

Given the relative benefits of each, it was decided that both types of rating scale would be included in the SJE. This mix of scales was designed to gain an indication of the candidate's knowledge of what action is appropriate in a given situation, and whether they would be likely to take that action. The SJE differentiated between scores for each scale, and the comparison of these scores allowed specific recommendations to be made for which type of questioning (situational, behavioural or both) should be used for each sub-criterion in the interview. Specifically, if the score on the helpful/unhelpful responses was lower than on the likely/ unlikely responses, this was regarded as an indicator that the participant would like to take the helpful action but cannot always spot which action is the helpful action and so a behavioural question was recommended. If the score pattern was vice-versa (likely / unlikely higher than helpful / unhelpful) a situational question was recommended because the pattern of responses indicates that the candidate can identify the 'helpful' action but for some reason doesn't always take this action.

However, it was subsequently decided that there should only be one type of rating scale – helpful/ unhelpful – for two reasons. Firstly, the analysis showed the helpful / unhelpful scale to have better validity. Secondly, feedback from assessors involved in the trials was that they found the variation in question types for each sub-criterion to be over complicated.

The use of one scale has simplified the recommendations made for the interview. Now, behavioural questioning is the default question type for all topic areas, with the addition of a situational question only recommended where the SJE sub-criterion score was very low or the consistency of responses was poor. In this instance the added benefit of the situational question is that it allows the interviewer to explore how the candidates' responses may depend on specific situations or circumstances. For more information on situational questioning please refer to the information provided below on the development of the MMI.

C.2 Multi-Modal Interview

C.2.1 Aims of the method

As outlined in Annex 3 Section 2.7, the Multi-Modal Interview (MMI) was developed to be used with the results of the SJE to make a judgement about the candidate's behavioural preferences in line with each of the selection criteria.

C.2.2 Development process and rationale

The design of the MMI aimed to address some of the limitations identified with the Criterion Based Interview (CBI). The CBI has been criticised as follows:

- The use of purely behavioural questioning.
- A method being used in isolation (ie not with an alternative measure of the same criteria).

As there was no existing interview schedule tailored to drivers that a) used a mix of question types and b) drew on results from a measure of behavioural preference, the MMI was developed in house by a team of Occupational Psychologists, with input from an expert psychometrician and the industry steering group.

Mix of question types

The interview uses two modes of questioning – behavioural (eg 'Give me an example of a time when…') and situational (eg 'What would you do if…').

These two question types are included in the design of the MMI for the following reasons:

- The two question types are measuring slightly different things. Behavioural questions measure past experiences and it is commonly thought that past behaviour is the best predictor of future behaviour.
- Situational questions provided the added benefit of measuring how the candidate would potentially behave when faced with a new experience, and exploring how behaviour may depend on the context. Meta-analyses have shown that both question types are associated with good validity. For example, Latham and Sue-Chan (1999) found a mean-corrected validity of .47 across 20 studies of situational interviews, and Taylor and Small (2002) found a mean-corrected validity of .45 across 30 situational interviews and .56 across 19 behavioural interviews.
- The inclusion of situational questions addresses the criticism of the CBI that it focuses on past experience rather than the potential of candidates who have not experienced particular situations (handling an emergency, for example).

Originally, the intention was that the MMI should also include motivational and biographical questioning so that it conformed to the underlying model of multi modal interviewing. However, it was felt this made the interview overlap too much with what happened in phases of the recruitment process that occur outside of the assessment centres (ie outside the remit of the project).

Combining with another measure

According to good practice, selection criteria should be measured using more than one method. The MMI complements the SJE as a measure of the behavioural criteria, making use of the SJE results to refine the interview process (ie to determine where additional situational questions are required).

The use of good practice in development

A structured interview

Research consistently demonstrates that structured interviews have higher validity and inter-rater reliability than unstructured interviews (eg Campion, Purcell and Brown, 1988). In line with this, the interview uses the same six topic areas for each candidate, standardised prompts and help words are provided for the interviewer to use, and the scoring of the interview is based upon the demonstration of behavioural indicators.

Behavioural indicators for scoring

As part of the development of the MMI, behavioural indicators were developed for each sub-criterion (see Figure 10 for an example). These indicators, which form the basis of the scoring of the MMI, were developed based on information collected from industry representatives with knowledge and / or experience of the train driver role. These representatives were asked to think of critical incidents in the driver role that help to differentiate a good driver from a poor driver.

The behavioural indicators were written to encapsulate the approach taken by good drivers in these critical incidents. The indicators were written to be as clear, measurable, relevant and user-friendly as possible, to aid the consistency of interviewers scoring each interview.

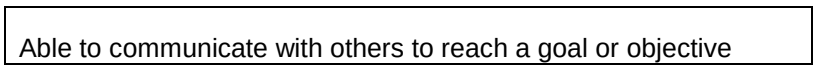


Figure 10 - An example behavioural indicator

The behavioural indicators have been refined based on feedback obtained from three sets of subject matter experts (SMEs); SMEs within RSSB with train driving experience, SMEs external to RSSB with train driving experience and SMEs external to RSSB with driver recruitment and selection experience.

The MMI topic areas and situational questions were reviewed by the SMEs with driver requirement and selection experience.

The increased structure within the interview should serve to reduce the information-processing demands upon the interviewer, which then enables them to make more accurate judgements of suitability

for the role (Gilbert, Krull, and Pelham, 1988; Gilbert and Krull, 1988).

Consideration of candidate impression management in the interview

A key concern when conducting an interview is that it does not allow the candidate to provide a false impression that they are better suited to the job than they really are. Levashina & Campion (2006) explain that faking in interviews is determined by the candidates' capacity, willingness and opportunity to fake.

With behavioural questions, candidates can potentially invent a past experience and use that as the basis for their interview answer. With situational questions, candidates can claim they would take certain actions that in reality they may not actually take.

In both instances, it is vital that the interviewer probes the candidate to determine how genuine the response actually is. This limits the candidates' opportunity to fake, by testing out their responses for sufficient detail and appropriate rationale. The structured nature of the interview also limits the opportunity to fake as the interviewer retains more control over the questioning process (Levashina & Campion, 2006).

C.2.3 Key iterations and refinements

A number of changes were made to the MMI following the trials in response to analysis findings and feedback from some interviewers that the MMI was over-complicated.

The changes made to the SJE following the trials (as explained earlier in this section) also have a follow-on impact on the MMI.

Changes to the SJE that impact the MMI

The trialled versions of the SJE had two types of scale – knowledge based and behavioural preference based. The relative scores on these scales were used to determine whether a situational or behavioural question was required for each sub-criterion. Now that only one rating scale is being used in the SJE, the MMI has been changed so that behavioural questions are used as default for each topic area, and a situational question is only added for the sub-criterion covered by that topic area if a) the candidate receives a low score on a sub-criterion or b) the candidate has poor consistency in his /her responses to covered by that sub-criterion.

Use of follow-up prompts

The use of follow-up prompts on the MMI was recommended for instances where the candidate a) received a low or inconsistent score for the sub-criterion covered by a topic area or b) the candidate was not providing enough evidence to determine a topic

area rating. In order to simplify the MMI, and allow each candidate the same opportunities within the interview, it was decided that follow-up prompts should be used as standard.

Removal of some behavioural indicators

At the end of the trial, the interviewers who had conducted the most interviews shared thoughts on what they thought worked well and not so well during the interview. It was agreed that some indicators were difficult to measure. In testing the criterion validity, analyses were tried with and without these indicators. Based on the findings, it was concluded that some indicators should be removed and others refined to make them more usable.

Changes to topic areas and scoring

The trialled version of the MMI consisted of eight topic areas, each relating to one or more sub-criteria. At the end of the interview, the interviewer allocated a score to each sub-criterion. The sub-criterion scores were then averaged to produce the three main criteria scores.

Some sub-scale scores were based on only a small number of indicators so, to enhance the reliability of the MMI the scoring was changed so that scores are allocated at the topic area rather than sub-criteria level, before being averaged to produce main criteria scores. This required some reorganisation of sub-criteria to topic areas. Using feedback from interviewers involved in the trial about which topic areas worked well, this enabled the number of topic areas to be reduced from eight to six to provide more efficient coverage of the 14 sub-criteria.

The bottom two points on the five-point rating scale were also redefined, so that 2 = none or some positive indicators, and one negative indicator, and 1 = none or some positive indicators and more than one negative indicators. In the trial, 2 = some of the positive indicators and one or more negative indicators, and 1 = no positive indicators, only negative indicators. The trial results showed that no participants scored a 1, and it was considered unlikely that any candidate applying for a job would give purely negative evidence. It is more meaningful to differentiate between how much negative evidence was provided (one indicator or more than one).

These changes have the added benefit of simplifying the interviewers' scoring task and reducing the 'cognitive load' for the interviewer.

D. Detailed demographic tables for the T948 validation study

Table 116 - Gender of participants who completed each trial method

	WAFV		VIGIL		2HAND		SJE		MMI		WCT v1		WCT v2		Group Bourdon		DTG		TRP		CBI	
Gender	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Male	111	91	112	92	112	92	60	87	81	88	59	86	65	89	129	94	153	93	162	93	150	93
Female	11	9	11	8	11	8	9	13	11	12	10	14	8	11	8	6	11	7	13	7	11	7
Total	**122**	**100**	**123**	**100**	**123**	**100**	**69**	**100**	**92**	**100**	**69**	**100**	**73**	**100**	**137**	**100**	**164**	**100**	**175**	**100**	**161**	**100**

Table 117 - Ethnicity of participants who completed each method in the trial

	WAFV		VIGIL		2HAND		SJE		MMI		WCT v1		WCT v2		Group Bourdon		DTG		TRP		CBI	
Ethnicity	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
White British	109	89	110	89	110	89	59	86	80	87	60	87	63	86	124	91	145	90	154	90	143	91
White other	4	3	4	3	4	3	2	3	2	2	2	3	2	3	4	3	8	4	8	5	8	5
Indian	1	1	1	1	1	1	1	1	3	3	1	1	2	3	3	3	2	1	3	2	2	1
Pakistani	2	1	2	2	2	2	2	3	1	1	1	1	1	1	1	1	1	1	2	1	1	1

Black Caribbean	3	3	3	3	3	3	4	6	5	5	4	6	3	4	0	1	1	1	1	1	1	1
Black African	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	2	2	1	2	1
Mixed race	3	3	3	3	3	3	1	1	1	1	1	1	2	3	1	1	1	1	2	1	1	1
Total	**122**	**100**	**123**	**100**	**123**	**100**	**69**	**100**	**92**	**100**	**69**	**100**	**73**	**100**	**134**	**100**	**161**	**100**	**172**	**100**	**158**	**100**

Table 118 – Age and experience demographic characteristics of the trial sample

Method	**Descriptive statistics**				
WAFV	*Mean*	*Min*	*Max*	*SD*	*N*
Age	41.30	25	61	7.80	122
Years on railway	9.02	0	38	9.02	121
VIGIL	Mean	Min	Max	SD	*N*
Age	41.37	25	61	7.81	123
Years on railway	10.04	0	38	9.23	123
2HAND	*Mean*	*Min*	*Max*	*SD*	*N*
Age	41.48	25	61	7.78	123
Years on railway	9.99	0	38	9.24	123

SJE version A	Mean	Min	Max	SD	N
Age	41.78	25	61	8.22	69
Years on railway	10.09	0	38	9.39	68
MMI	Mean	Min	Max	SD	N
Age	41.30	26	58	7.20	92
Years on railway	10.57	0	33	8.20	91
WCT v1	Mean	Min	Max	SD	N
Age	41.54	25	59	7.36	69
Years on railway	10.01	0	32	8.32	67
WCT v2	Mean	Min	Max	SD	N
Age	41.99	26	61	8.29	73
Years on railway	10.12	0	38	9.62	72
Group Bourdon	Mean	Min	Max	SD	N
Age	36.36	21	62	7.68	137
Years on railway	1.740	0	14	2.68	124
DTG	Mean	Min	Max	SD	N

Age	36.67	21	62	7.55	164
Years on railway	1.95	0	14	1.20	152
TRP	Mean	Min	Max	SD	N
Age	36.86	21	62	7.7	175
Years on railway	1.90	0	14	2.55	162
CBI	Mean	Min	Max	SD	N
Age	36.60	21	62	7.55	161
Years on railway	1.95	0	14	2.50	149

Table 119 – Role of participants who completed each trial method

	WAFV		VIGIL		2HAND		SJE		MMI		WCT v1		WCT v2		Group Bourdon		DTG		TRP		CBI	
Role	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Train drivers	91	75	91	74	91	74	48	70	76	83	51	75	54	75	42	30	70	43	71	40	68	44
Trainee drivers	16	13	16	13	16	13	10	15	8	9	8	12	8	11	79	58	86	53	87	50	86	56
Driver managers	3	2	3	2	3	2	2	3	1	1	2	3	2	3	1	1	1	1	1	1	1	1
Failed candidates	7	6	7	6	7	6	4	6	1	1	2	3	5	7	7	5	0	0	7	4	0	0
Other	5	4	6	5	6	5	4	6	4	4	5	7	3	4	8	6	7	3	9	5	0	0
Total	**122**	**100**	**123**	**100**	**123**	**100**	**68**	**100**	**90**	**100**	**68**	**100**	**72**	**100**	**137**	**100**	**164**	**100**	**175**	**100**	**155**	**100**

Table 120 – Educational level of participants who completed each trial method

	WAFV		VIGIL		2HAND		SJE		MMI		WCT v1		WCT v2		Group Bourdon		DTG		TRP		CBI	
Educational level	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
No formal qualifications	10	8	10	8	10	8	8	12	7	8	5	7	8	11	5	4	7	5	7	5	6	4
GCSEs	43	35	43	35	43	35	18	27	25	28	26	38	18	25	47	39	59	39	59	39	57	38
GNVQs	31	26	31	25	31	25	17	25	22	24	16	24	19	26	20	16	22	14	22	14	22	15
A-levels / Scottish Highers	28	23	29	25	29	25	15	22	28	31	17	25	20	28	36	28	41	27	41	27	41	28
Degree	9	8	9	7	9	7	9	13	9	10	4	6	7	10	17	13	23	15	23	15	23	15
Total	**121**	**100**	**122**	**100**	**122**	**100**	**68**	**100**	**91**	**100**	**68**	**100**	**72**	**100**	**125**	**100**	**152**	**100**	**152**	**100**	**149**	**100**

Table 121 - Participants with / without dyslexia who completed each trial method

	WAFV		VIGIL		2HAND		SJE		MMI		WCT v1		WCT v2		Group Bourdon		DTG		TRP		CBI	
Dyslexia	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Yes	4	3.3	4	3.3	4	3.3	1	2	3	3	1	1	3	4	3	7.7	2	3.8	3	4.8	2	4
No	116	96.7	117	96.7	117	96.7	66	98	87	97	66	99	69	96	36	92.3	51	96.2	59		51	96
Total	**120**	**100**	**121**	**100**	**121**	**100**	**67**	**100**	**90**	**100**	**67**	**100**	**72**	**100**	**39**	**100**	**53**	**100**	**62**	**100**	**53**	**100**

Table 122 – First language of participants who completed each trial method

	WAFV		VIGIL		2HAND		SJE		MMI		WCT v1		WCT v2		Group Bourdon		DTG		TRP		CBI	
Language	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
English	117	96	118	96	118	96	68	99	87	96	65	96	71	97	121	94.5	148	95.5	158	95.2	52	96
Other	4	4	4	4	4	4	1	1	4	4	3	4	2	3	7	5.5	7	4.5	8	4.8	2	4
Total	**121**	**100**	**122**	**100**	**122**	**100**	**69**	**100**	**91**	**100**	**68**	**100**	**73**	**100**	**128**	**100**	**155**	**100**	**166**	**100**	**54**	**100**

E. Study variables

E.1 Job performance variables

The table below sets out a rating scale for 11 indicators describing operational driving performance ranging from **1 (very poor) to 5 (excellent).** The driver managers were asked to try to bear in mind the differences between all the drivers / trainees they have known.

Table 123 – Operational driving performance data

Performance indicators	Rating scale				
	1 (very poor)	**2**	**3**	**4**	**5 (excellent)**
Overall train handling (Acceleration and braking technique / Speed and distance judgment)	Regular significant errors	Occasional significant errors	Regular minor errors	Occasional minor errors	Perfect record
Application of rules	Significant non-compliance with the relevant sections of the rulebook involving safety consequences	Significant non-compliance with the relevant sections of the rulebook involving no safety consequences	Regular very minor non-compliance with the relevant sections of the rulebook	Occasional very minor non-compliance with the relevant sections of the rulebook	Perfect record of applying the relevant sections of the rulebook
SPAD record	2 or more (Cat A) SPADs	1 (Cat A) SPAD	No SPADs	No SPADs	No SPADs
Collisions or derailments (inc. buffers)	2 or more collisions or derailments involving reportable injuries and / or material damage	1 collision or derailment involving reportable injuries and / or material damage	1 collision or derailment involving no reportable injuries or material damage	A near miss collision or derailment involving no reportable injuries or material damage	No occurrences

Performance indicators	Rating scale				
	1 (very poor)	**2**	**3**	**4**	**5 (excellent)**
Speeding record	3 or more occurrences of speeds greater than 10mph in excess	1 or 2 occurrences of speeds greater than 10mph in excess	1 or 2 occurrences of speed between 6 & 10mph in excess	1 occurrence of speed up to 5mph in excess	No occurrences
Station overrun record	3 or more overruns	2 overruns	1 overrun	1 missed car stopping mark but still in platform	No occurrences
Station disregard record	3 or more disregards	2 disregards	1 disregard	1 near miss disregard but still in platform (eg heavy braking)	No occurrences
Operation / isolation of safety systems	2 or more unauthorised operations / isolations	1 unauthorised operations / isolations	No unauthorised occurrences	No unauthorised occurrences	No unauthorised occurrences
Preparation, disposal and handover of trains	4 or more omissions/ incorrect applications	3 omissions/ incorrect applications	2 omissions/ incorrect applications	1 omission/ incorrect application	No omissions/ incorrect applications
Workplace formal assessment (either practical or lack of underpinning knowledge)	Has required reassessment for 3 or more units	Has required reassessment for 2 units	Has required reassessment for 1 unit	Good record of formal assessment - has not required reassessment	Perfect record of formal assessment against the Competence Standards for Train Drivers

Performance indicators	Rating scale				
	1 (very poor)	2	3	4	5 (excellent)
Handling abnormal events / calmness under pressure (if not applicable, leave blank)	When exposed to an abnormal event, has rarely handled it appropriately	When exposed to an abnormal event, has occasionally handled it appropriately	When exposed to an abnormal event, has generally handled it appropriately	When exposed to an abnormal event, has almost always handled it appropriately	When exposed to an abnormal event, has always handled it appropriately

The table below sets out the rating scales for three performance indicators which cover how well a trainee performed during their initial training period. The ratings describe performance ranging from **very poor (1) to excellent (5)**. The training managers were asked, for each trainee driver that has participated in the trial, to enter the performance rating, from 1 to 5, that best describes the performance of that trainee against each indicator.

Table 124 – Trainee driver training performance

Performance indicators	Rating scale				
	1 (very poor)	2	3	5	5 (excellent)
Rules assessment (understanding, retention and application)	The trainee was slower than most to develop and retain knowledge of RULES and needed help to apply the underlying concepts to real world settings.		The trainee was able to develop and retain knowledge of RULES within a reasonable timeframe but needed some help to apply the underlying concepts to real world settings.		The trainee was quick to develop and retain knowledge of RULES and to show the ability to effectively apply the underlying concepts to real world settings with minimal help.

Performance indicators	Rating scale				
	1 (very poor)	2	3	5	5 (excellent)
Traction - theoretical (understanding, retention and application)	The trainee was slower than most to develop and retain knowledge of TRACTION and needed help to apply the underlying concepts to real world settings.		The trainee was able to develop and retain knowledge of TRACTION within a reasonable timeframe but needed some help to apply the underlying concepts to real world settings.		The trainee was quick to develop and retain knowledge of TRACTION and to show the ability to effectively apply the underlying concepts to real world settings with minimal help.
Traction – practical handling (maintaining control of the vehicle/ fault identification and diagnosis/ coupling and decoupling/ braking/ operating cab-display and interface equipment/ operating physical features of the cab and carriages)	The trainee was slower than most to demonstrate the ability to perform PRACTICAL TRACTION tasks and needed more help than usual to do so.		The trainee demonstrated the ability to perform PRACTICAL TRACTION tasks without fault within a reasonable timeframe and after receiving some help.		The trainee quickly demonstrated the ability to perform PRACTICAL TRACTION tasks without fault and with no additional help.

Table 125 – Manager's ratings of performance

Job performance data category	Data variable	Rating scale and further information
Manager ratings of communication	Speaking	Each data variable was defined using behavioural markers. Managers rated on a 1-5 scale (1 = could be better in all of these things to 5 = shows all these things to a high level. Is one of the best I know at this).
	Reading	
	Listening	
	Written communication - Accuracy Written communication - Inclusion of key details Written communication - Legibility Written communication - Logical order Written communication - Conciseness Written communication - Overall	Each data variable was defined using behavioural markers. Managers rated on a 1-5 scale (1 = could be better in all of these things to 5 = shows all these things to a high level. 'Is one of the best I know at this'). An overall written communication score was created by totalling the written communication section scores. This was so that manager's rating of written communication could be correlated with the overall score on the WCT.
	Overall communication	An overall communication score was created by averaging overall written communication, speaking, reading and listening. Data was also collected in previous trial in relation to this (managers rated on a 1-5 scale where each scale point was defined in terms of how often the candidate provides relevant information to the right people, clearly and precisely, both written and verbal, at the appropriate times). The data from both trials was consolidated, with preference given to the current trial data where both were available.
Manager ratings of behaviour	Dependability	Each data variable was defined using behavioural markers. Managers rated on a 1-5 scale (1 = could be better in all of these things to 5 = shows all these things to a high level. Is one of the best I know at this).
	Commitment to work	
	Attention to detail	
	Ability to check and not make assumptions	
	Compliance with rules and procedures	

Job performance data category	Data variable	Rating scale and further information
	Proactivity	
	Tenacity	
	Assertiveness	
	Calmness under pressure	
	Reactivity to stress	
	Social need	
	Sensation seeking	
	Need for external stimulation	
	Attitude to work and people	Ratings collected as above in current trial. Data was also collected in previous trial in relation to this (managers rated on a 1-5 scale where each scale point was defined in terms of how often the candidate displays a conscientious and dependable attitude to work such as never missing on duty, late arrival to service or late starts). The data from both trials was consolidated, with preference given to the current trial data where both were available.
	Conscientiousness	A score attained by averaging all the relevant sub-criteria (dependability, consolidated attitude to work and people, commitment to work, attention to detail, checking and rule compliance).
	Dealing with challenging situations	A score attained by averaging all the relevant sub-criteria (proactivity, tenacity, assertiveness, calmness under pressure, reactivity to stress).
	Tolerance for low stimulation	A score attained by averaging all the relevant sub-criteria (social need, sensation seeking, need for stimulation).

E.2 Assessment method scores

Some assessment methods have norm groups although the methods developed in-house do not. Where norm group data is available, these have also been used to explain what good performance means.

E.2.1 WAFV

Table 126 – WAFV test scores

Test score	Description	High RAW score means good performance	High NORM scores mean good performance
Number of missed reactions	This is the number of stimuli to which no response was made within 1500 ms. A high test score leads to a low percentile rank and indicates problems in the continuous maintaining of vigilance. It is particularly important to note whether the omissions are evenly distributed or whether they occur in clusters (as would be expected, for example, if the subject is experiencing some form of micro-sleep). Alternatively, omissions may become more frequent towards the end of the test (comparison between first and second halves of the test), pointing to a possible problem in maintaining vigilance.	No	Yes
Mean reaction time	This variable is a logarithmic mean of the individual reaction times. The advantage of using a logarithmic mean is that it takes account of the expected skew of the distribution of the reaction times. A high test score leads to a low percentile rank and indicates a slow processing speed in identifying the stimuli in the vigilance tasks.	No	Yes
Number of false alarms	This is the number of times a reaction key was pressed in response to irrelevant stimuli or when no stimulus had been	No	Yes

	presented. Raised error rates should normally be interpreted as an indication not of an impairment of the intensity of attention but (at least partially) of an impairment of selectivity, unless they vary systematically over time.		

E.2.2 VIGIL

Table 127 – VIGIL test scores

Test score	Description	High RAW score means good performance	High NORM scores mean good performance
Number of correct	The total number of correct reactions to critical stimuli. A 'correct' reaction is a reaction to a double jump such that the response button is pressed before the next jump takes place. This variable describes the accuracy of the respondent's observation over the test as a whole. Individuals with a high score (PR>84) on this variable are considered to have excellent visual observation ability and are good at noticing detail in monotonous stimulus situations. These respondents are good at identifying the double jumps and are able to sustain their attention over a relatively lengthy period of time.	Yes	Yes
Number of incorrect	The total number of incorrect reactions. A reaction is classed as 'incorrect' if it is made when no critical stimulus has occurred ('illusion errors' or 'false alarms'). Reactions in the absence of a critical stimulus occur infrequently in the existing comparison and norm samples. This variable is therefore a control variable; it is reported in order to verify that the respondent has understood the instructions and is taking the task seriously.	No	Yes

Mean value of reaction time correct (sec)	This variable measures the average time that elapses between presentation of a critical stimulus and correct pressing of the button. It provides information about the respondent's speed of information processing in vigilance situations and also takes account of aspects of his or her motor reaction ability. Respondents with a high percentile rank (PR>84) on this variable are therefore very good at perceiving and reacting appropriately to suddenly occurring stimuli even in long-lasting, monotonous situations. They are able to sustain their attention over a relatively lengthy period of time even under monotonous conditions.	No	Yes

E.2.3 2HAND

Table 128 – 2HAND test scores

Test score	Description	High RAW score means good performance	High NORM scores mean good performance
Total mean duration	The variable 'total mean duration' corresponds to the average time needed to run through the track. It measures the ability respond to the features of the track (simple, difficult) and use appropriate fine motor movements. Moreover, a certain readiness to take risks is also measured.	No	No
Total mean error duration	The total error duration is the time during which the cursor – calculated over all runs – was outside of the range of tolerance of the established track area. It shows how well the respondent is able to turn small deviations from the intended route into the appropriate compensatory motions. This measurement therefore includes, in addition to exactness of fine motor movements, accuracy of information processing.	No	No
Total percent error duration	This is defined by the ratio of total error duration to total duration.	No	No
Coordination difficulty	This is a measure of the coordination performance of the respondent. Numerically, coordination difficulty is defined by the time-ratio required to master an equally long track with or without coordination.	No	No

E.2.4 Group Bourdon

Table 129 – Group Bourdon test scores

Test score	Description	High RAW score means good performance
Total productions	The test requires applicants to identify and mark examples of a particular pattern of dots which are embedded in a set of other dot patterns. Total productions is the total number of stimuli that are checked within the time limit.	Yes
Total omissions	Total omissions are the total number of correct patterns of dots which the candidate failed to mark.	No
Total faults	Total faults are the total number of errors that the candidate made, where the incorrect pattern of dots were marked.	No

E.2.5 DTG

Table 130 - Determinations Gerat (DTG) test scores

Test score	Description	High RAW score means good performance
Part 3 good scores	This is the number of correct responses made in the timed phase of the test.	Yes
Part 3 omissions	This is the number of items where no response is made during the timed phase of the test.	No
Self-paced wrong	This is the number of items where errors are made during the self-paced phase.	No

E.2.6 TRP1

Table 131 – Trainability for Rules and Procedures TRP1 test score

Test score	Description	High RAW score means good performance
TRP1	Candidates listen to a recording of fictitious safety-related rules information. They learn the information and then answer from memory, 18 questions based on the information they have just heard and read. TRP1 is a measure of verbal memory and comprehension.	Yes

E.2.7 TRP2

Table 132 – Trainability for Rules and Procedures TRP2 test score

Test score	Description	High RAW score means good performance
TRP2	Candidates are presented with fictitious dials for a train cab and rules relating to their functioning. Candidates are required to apply the rules to decide the order in which the dials should be checked. There are 43 sets of dials for candidates to work through. TRP2 measures verbal comprehension, rule application and non-verbal reasoning.	Yes

E.2.8 SJE

Table 133 - SJE Scores

Test score	Description	High RAW / z100 score means good performance
Conscientiousness	Comprises of an average of the following sub-criteria scores (which are used in the MMI): Dependability Attitude to work and people Commitment to work Attention to detail Ability to check and not make assumptions	Yes

	Compliance with rules and procedures The raw main criteria score is banded into 'low' or 'moderate / good' based on z100 score calculations from the trial sample. The SJE band is combined with the MMI score to make an overall pass/fail decision for that criterion.	
Dealing with challenging situations (DCS)	Comprises of an average of the following sub-criteria scores (which are used in the MMI): Proactivity Tenacity Assertiveness Calmness under pressure Reactivity to stress The main criteria scores are banded as described under 'conscientiousness' above.	Yes
Tolerance for low stimulation (TLS)	Comprises of an average of the following sub-criteria scores (which are used in the MMI): Social need Sensation seeking Need for external stimulation The main criteria scores are banded as described under 'conscientiousness' above.	Yes
Sub-criteria scores	Sub-criteria scores are banded into 'low', 'moderate' or 'good' based on z100 scores from the trial sample ('low' = z100 < 77.5, 'moderate' = ≥77.5 z100 < 90.5, 'good' = z100 ≥ 90.5) . These scores are used to inform the MMI. If a sub-criterion is 'low' or 'moderate' or has 'weak' consistency then a situational question is required in the MMI to cover that sub-criterion.	Yes
Sub-criterion score consistency	A consistency score (weak, medium, strong) is generated for each sub-criterion based on established data limits (these vary for each sub-criterion depending on the strength and number of SJE actions weighted to the sub-criterion) for the purposes of informing the MMI. If a sub-criterion is 'low' or 'moderate' or has 'weak' consistency then a situational question is required in the MMI to cover that sub-criterion.	Yes

E.2.9 MMI

Table 134 - MMI Scores

Test score	Description	High RAW score means good performance
Conscientiousness	Ratings are applied to topic areas based on the behavioural indicators as follows: 5 = All of the positive indicators and no negative indicators 4 = Majority of the positive indicators and no negative indicators 3 = Less than half of the positive indicators and no negative indicators 2 = Some or none of the positive indicators and one negative indicator 1 = Some or none of the positive indicators and more than one negative indicator This score is an average of Topic area 1, Topic area 2, and Topic area 3* and should be interpreted as follows: 5 = Exemplary, all positive indicators 4 = Very good, majority of positive indicators 3 = Acceptable, less than half of the positive indicators FAIL (one or more negative indicators)	Yes
Dealing with challenging situations (DCS)	Ratings are applied to topic areas in the same way as described above. Comprises of an average of Topic area 4 and Topic area 5* which is then banded FAIL / 3 / 4/ 5 as described above.	Yes
Tolerance for low stimulation (TLS)	Ratings are applied to the topic area in the same way as described above. Topic area 6 score* which is then banded FAIL / 3 / 4/ 5 as described above.	Yes
Verbal communication	A five point rating scale is used as follows: 5= All positive indicators demonstrated	Yes

	to an exemplary standard 4 = All positive indicators demonstrated to a good standard 3 = Most positive indicators are displayed to an acceptable level, training should improve the one or two minor instances of negative communication evidence 2 = Some substantial examples of one or more negative indicators / difficult to understand in parts 1 = Significant negative indicators / difficult to understand in the majority of the interview	

* Note that this is a post-trial amendment, in the trial the main criteria scores were calculated from relevant sub-criteria scores.

E.2.10 CBI

The higher the score, the better it is.

Table 135 - CBI Scores

Test score	Description	High RAW score means good performance
Follows set rules and procedures	These scores were extracted from the RACF database. A variety of score coding had been used and so CBI scores were recoded as follows for the purposes of the analysis: A/ very good pass/ good pass = 3, B/ borderline pass / pass = 2, C/D/fail = 1. These score bands are based upon positive evidence as defined by behavioural indicators.	Yes
Conscientiously works to meet training and job demands		
Remain calm in emergency and stressful situations		
Proactive and tenacious		
Can spend time alone and does so effectively		
Ability to communicate effectively verbally and in writing		

E.2.11 WCT

Table 136 - WCT Scores

Test score	Description	High RAW score means good performance
Instead of being used as a pass/ fail measure, it is recommended that the WCT provides a qualitative assessment of written communication indicating potential development needs during training.		
Legibility score	Comprises of a total score out of 2 points that is marked using a scoring framework. This score is banded 'low', 'moderate' or 'good' accordingly. Candidates who do not meet ≥ 1 (do not have legible handwriting) would be unable to obtain an overall moderate or good score on the WCT.	Yes
Accuracy	Comprises of a total score of 6 points that is marked using a scoring framework. This score is banded 'low', 'moderate' or 'good' accordingly.	Yes
Written comprehension score	Comprises of a total score of 10 points that is marked using a scoring framework. This score is banded 'low', 'moderate' or 'good' accordingly.	Yes
Structure score	A total score of 4 comprises of up to 2 points that can be awarded for logical sequence and up to 2 points that can be awarded for relevance. The total score is marked using a scoring framework and is banded 'low', 'moderate' or 'good' accordingly.	Yes
Overall WCT score	The maximum total number of points available is 22 (the sum product of the maximum scores available from each WCT section).	Yes
Overall WCT band	On its own, the overall score does not indicate the level of written communication skill that a candidate has. Overall banding is as important as it indicates the quality of written communication and potential training requirements. The bands are split as follows: • 'Good' = candidate achieves good on all sections. It is unlikely to require special attention during communication training. The 'good' band has been set at the minimum levels reached by almost every candidate in these trails. This is the only band where a candidate's total score will indicate the quality of written comprehension they have. • 'Moderate' = candidate achieves moderate scores on one or more of the WCT sections. This candidate	Yes

	may require some attention during communication training compared to other candidates. The areas of training are indicated by the areas highlighted as moderately scored. The 'moderate' band has been set at a level to reflect a score that is close to good but with some weakness. • 'Low' = candidate achieves a low score on one or more of the WCT sections. This candidate may require extensive training in certain aspects of written communication than others (as highlighted by the qualitative comments) compared to other candidates. The 'low' band was set at two standard deviations below the mean to reflect the position of very poor performance. • 'Illegible' = candidate scores low on legibility, regardless of the score on other sections. This is not likely other than in special circumstances so should be explored further in conversation with the candidate to understand any underlying issues. The 'illegible' band of the overall score has been set using the rationale that in order for a candidate to have a basic level of written communication their writing must be legible to others.	

F. Expected relationships between job performance measures and assessment method scores

Tables of the expected relationships between psychometric assessment scores and job performance variables are presented below.

F.1 WAFV

Table 137 - WAFV Predicted relationships table

WAFV	Operational performance data												Training performance data		
	Train handling	*Application of rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Workplace formal assessment*	*Handling abnormal events*	*Rules*	*Traction theory*	*Traction practical*
Number of missed reactions	+		+	+	+	+									
Mean reaction time	+		+	+	+	+									
Number of false alarms	+		+	+	+	+									

F.2 VIGIL

Table 138 - VIGIL Predicted relationships table

VIGIL	Operational performance data												Training performance data		
	Train handling	*Application of rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Workplace formal assessment*	*Handling abnormal events*	*Rules*	*Traction theory*	*Traction practical*
Number of correct	+		+	+	+	+									
Number of incorrect	+		+	+	+	+									
Mean value of reaction time correct (sec)	+		+	+	+	+									

F.3 2HAND

Table 139 – 2HAND Predicted relationships table

2HAND	Operational performance data													Training performance data		
	Train handling	*Application of rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Workplace formal assessment*	*Handling abnormal events*	*Train handling during abnormal situations*	*Rules*	*Traction theory*	*Traction practical*
Total mean duration	+												+			+

2HAND	Operational performance data													Training performance data		
	Train handling	*Application of rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Workplace formal assessment*	*Handling abnormal events*	*Train handling during abnormal situations*	*Rules*	*Traction theory*	*Traction practical*
Total mean error duration	+												+			+
Total percent error duration	+												+			+
Coordination difficulty	+												+			+

F.4 Group Bourdon

Table 140 – Group Bourdon Predicted relationships table

Group Bourdon	Operational performance data												Training performance data		
	Train handling	*Application of rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Workplace formal assessment*	*Handling abnormal events*	*Rules*	*Traction theory*	*Traction practical*
Total productions	+		+	+	+	+	+	+	+	+			+	+	+
Total omissions	+		+	+	+	+	+	+	+	+			+	+	+
Total faults	+		+	+	+	+	+	+	+	+			+	+	+

F.5 DTG

Table 141 - DTG Predicted relationships table

DTG	Operational performance data												Training performance data		
	Train handling	*Application of rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Workplace formal assessment*	*Handling abnormal events*	*Rules*	*Traction theory*	*Traction practical*
Part 3 good scores	+		+	+	+	+	+	+	+						+
Part 3 omissions	+		+	+	+	+	+	+	+						+
Self-paced wrong	+		+	+	+	+	+	+	+						+

F.6 TRP1

Table 142 – TRP1 Predicted relationships table

TRP1	Operational performance data												Training performance data		
	Train handling	*Application of rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Workplace formal assessment*	*Handling abnormal events*	*Rules*	*Traction theory*	*Traction practical*
TRP1		**+**								+	+		+	+	+

F.7 TRP2

Table 143 - 2HAND Predicted relationships table

TRP2	Operational performance data												Training performance data		
	Train handling	*Application of rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Workplace formal assessment*	*Handling abnormal events*	*Rules*	*Traction theory*	*Traction practical*
TRP2	+	**+**	+	+	+	+	+	+	+	+		+	+	+	+

F.8 WCT

Table 144 - Predicted relationships between WCT and job performance factors

	Job performance factor (training)		
	Rules assessment	*Traction theoretical*	*Traction practical*
WCT: Overall WCT score	**+**	**+**	+

Table 145 - Predicted relationships between WCT and behavioural performance factors

WCT total scores	Manager's performance rating on written communication					
	Accuracy	*Details*	*Legibility*	*Logical*	*Conciseness*	*Total*
Detail section	+					+
Summary section		+				+
Legibility section			+			+
Structure - logical sequencing				+		+
Structure – concise					+	+
Overall WCT score	**+**	**+**	+	**+**	**+**	+

F.9 SJE

Table 146 – SJE Predicted relationships table

SJE	Operational performance data												Training performance data			Comm	Managers' behaviour ratings		
	Train handling	*Rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Formal assessment*	*Abnormal events*	*Rules*	*Traction theory*	*Traction practical*		*Consc.*	*DCS*	*TLS*
Conscientiousness	+	**+**	**+**	+		+	+		+	+			**+**	+	+	**+**	**+**		
DCS									+	+						**+**		+	
TLS	+	+	+	+	+		+	+											+

F.10 MMI

Table 147 – MMI Predicted relationships table

MMI	Operational performance data												Training performance data			Comm	Managers' behaviour ratings		
	Train handling	*Rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Formal assessment*	*Abnormal events*	*Rules*	*Traction theory*	*Traction practical*		*Consc.*	*Challenging situations*	*Tolerance*
Conscientiousness	+	**+**	**+**	+		+	+		+	+			**+**	+	+		**+**	**+**	
DCS									+	+							**+**		+
TLS	+		+	+	+		+	+											
Verbal communication																**+**			

F.11 CBI

Table 148 – CBI Predicted relationships table

CBI	Operational performance data											Training performance data			Comm	Managers' behaviour ratings		
	Train handling	*Rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Formal assessment*	*Abnormal events*	*Rules*	*Traction theory*	*Traction practical*		*Consc.*	*Challenging situations*	*Tolerance*
Rules	+	**+**	**+**	+		+	+		+	+			**+**	+	+		**+**		
Conscientiousness	+	**+**	**+**	+		+	+		+	+			**+**	+	+		**+**		
Emergency situations									+			+						**+**	

CBI	Operational performance data												Training performance data			Comm	Managers' behaviour ratings		
	Train handling	*Rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Formal assessment*	*Abnormal events*	*Rules*	*Traction theory*	*Traction practical*		*Consc.*	*Challenging situations*	*Tolerance*
Proactive and tenacious									+			+						**+**	
TLS	+		+	+	+		+	+											+
Communication																**+**			

G. Full correlation tables for T948 validation study

Results for all correlations tested according to the expected relationships.

G.1 WAFV

Table 149 - Correlations between WAFV scores and performance data

WAFV		Operational performance data												Training performance data		
		Train handling	*Application of rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Workplace formal assessment*	*Handling abnormal events*	*Rules*	*Traction theory*	*Traction practical*
Number of missed reactions	*r*	**-.21**		**-.19**	.06	.09	-.16									
	p	.03		.05	.29	.21	.07									
	N	80		80	80	80	80									
Mean reaction time	*r*	-.13		**-.19**	.11	.14	-.06									
	p	.13		.05	.16	.10	.31									
	N	80		80	80	80	80									
Number of false alarms	*r*	**-.36**		**-.20**	.05	.10	**-.30**									
	p	<.001		.03	.34	.19	.00									
	N	80		80	80	80	80									

G.2 VIGIL

Table 150 - Correlations between VIGIL scores and performance data

VIGIL		Operational performance data												Training performance data		
		Train handling	*Application of rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Workplace formal assessment*	*Handling abnormal events*	*Rules*	*Traction theory*	*Traction practical*
Number of correct	*r*	-.05		-.10	.10	.03	-.02									
	p	.32		.20	.20	.39	.42									
	N	81		81	81	81	81									
Number of incorrect	*r*	.03		.07	-.06	.07	-.03									
	p	.41		.27	.30	.28	.41									
	N	81		81	81	81	81									
Mean value of reaction time correct (sec)	*r*	-.05		-.10	.10	.03	-.02									
	p	.32		.20	.20	.39	.41									
	N	80		80	80	80	80									

G.3 2HAND

Table 151 - Correlations between 2HAND scores and performance data

2HAND		Operational performance data													Training performance data		
		Train handling	*Application of rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Workplace formal assessment*	*Handling abnormal events*	*Train handling during abnormal situations*	*Rules*	*Traction theory*	*Traction practical*
Total mean duration	*r*	**-.28**												**-.32**			-.22
	p	.03												.02			.18
	N	81												80			34
Total mean error duration	*r*	-.02												**.05**			.05
	p	.46												.04			.42
	N	81												80			34
Total percent error duration	*r*	.14												.17			.06
	p	.21												.17			.41
	N	81												80			34
Coordination difficulty	*r*	-.01												-.04			-.12
	p	.45												.35			.25
	N	81												80			34

G.4 Group Bourdon (paper version)

Table 152 - Correlations between Group Bourdon (paper version) scores and performance data

Group Bourdon (paper version)		Operational performance data												Training performance data		
		Train handling	*Application of rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Workplace formal assessment*	*Handling abnormal events*	*Rules*	*Traction theory*	*Traction practical*
Total productions	*r*	.19		.*	.*	.22	**.32**	**-.37**	.02	.24	-.12			.25	-.06	.07
	p	.13		.00	.00	.10	.03	.01	.47	.08	.28			.17	.41	.40
	N	38		38	38	38	38	38	38	38	38			17	17	17
Total omissions	*r*	.21		*	*	.12	.16	.09	.03	.12	.15			**.46**	**.45**	**.57**
	p	.11		.00	.00	.23	.17	.30	.17	.24	.18			.03	.03	.01
	N	38		38	38	38	38	38	38	38	38			17	17	17
Total faults	*r*	.10		*	*	.10	-.05	.05	.05	-.05	.22			-.10	.03	-.12
	p	.28		.00	.00	.28	.39	.38	.40	.39	.09			.35	.45	.33
	N	38		38	38	38	38	38	38	38	38			17	17	17

* Cannot be computed because at least one of the variables is constant

G.5 DTG

Table 153 - Correlations between DTG scores and performance data

DTG		Operational performance data												Training performance data		
		Train handling	*Application of rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Workplace formal assessment*	*Handling abnormal events*	*Rules*	*Traction theory*	*Traction practical*
Part 3 good responses	*r*	.03		.08	-.08	.04	-.09	-.13	.05	.06						.08
	p	.42		.27	.27	.38	.24	.16	.35	.31						.36
	N	64		64	64	64	64	64	64	63						25
Part 3 omissions	*r*	-.05		.03	-.03	-.02	.05	.08	.05	-.02						-.08
	p	.36		.40	.40	.45	.36	.26	.34	.43						.35
	N	64		64	64	64	64	64	64	63						25
Self-paced wrong	*r*	.54		*	*	.16	.09	-.27	.35	*						-.01
	p	.06		.00	.00	.33	.40	.23	.16	.00						.49
	N	10		10	10	10	10	10	10	9						19

* Cannot be computed because at least one of the variables is constant

G.6 TRP1

Table 154 - Correlations between TRP1 scores and performance data

TRP1		Operational performance data												Training performance data		
		Train handling	*Application of rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Workplace formal assessment*	*Handling abnormal events*	*Rules*	*Traction theory*	*Traction practical*
TRP1	*r*		-.08								-.07	-.03		-.15	-.30	**-.48**
	p		.27								.28	.41		.24	.07	<.01
	N		65								65	65		25	25	25

G.7 TRP2

Table 155 - Correlations between TRP2 scores and performance data

TRP2		Operational performance data												Training performance data		
		Train handling	*Application of rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Workplace formal assessment*	*Handling abnormal events*	*Rules*	*Traction theory*	*Traction practical*
TRP2	*r*	.01	-.05	-.02	.02	-.01	-.13	.00	.09	.20	-.10		-.12	-.06	-.20	**-.46**
	p	.49	.36	.44	.44	.45	.15	.50	.24	.06	.21		.17	.38	.17	.01
	N	65	65	65	65	65	65	65	65	64	65		65	25	25	25

G.8 WCT

Table 156 - Pearson correlations between WCT and job performance factors (original versions)

Overall WCT score		Job performance factor (training)		
		Rules assessment	*Rules assessment*	*Rules assessment*
Version 1	*r*	.52	.49	.34
	P	.00	.00	.06
	N	23	23	23
Version 2	*r*	.38	-.04	.35
	P	.06	.44	.08
	N	18	18	18

Table 157 - Pearson correlations between WCT (original versions) and behavioural performance factors

WCT Version 1 total scores		Manager's performance rating on written communication					
		Accuracy	*Details*	*Legibility*	*Logical*	*Conciseness*	*Total*
Detail section	*r*	.21					.37
	p	.06					.00
	N	61					61
Summary section	*r*		.01				.07
	P		.47				.29
	N		61				61
Legibility section	*R*			-.03			.0
	p			.41			.37
	N			61			61
Structure – logical sequencing	*r*				-.06		.07
	p				.33		.29
	N				61		61
Structure – concise	*r*					.07	-.06
	p					.31	.32
	N					61	61
Overall WCT score	*r*	.12	.08	.39	.10	.19	.22
	p	.17	.26	.00	.23	.08	.04
	N	61	61	61	61	61	61

Table 158 - Pearson correlations between WCT (original versions) and job performance factors

WCT version 2		Manager's performance rating on written communication					
		Accuracy	*Details*	*Legibility*	*Logical*	*Conciseness*	*Total*
Detail section	*r*	.13					.14
	p	.17					.14
	N	62					62
Summary section	*r*		.04				.12
	p		.38				.18
	N		62				62
Legibility section	*r*			.41			.35
	p			.00			.00
	N			62			62
Structure – logical sequence	*r*				-.05		-.15
	p				.35		.13
	N				62		62
Structure – concise	*r*					.33	.27
	p					01	.02
	N					62	62
Overall WCT score	*r*	.15	.12	.16	.31	.22	.22
	p	.13	.17	.12	.01	.04	.05
	N	62	62	62	62	62	62

G.9 SJE

Table 159 - Correlations between SJE version A main criteria scores and performance data

SJE Measure		Operational performance data												Training performance data			Communication
		Train handling	*Rules*	*SPAD record*	*SPAD risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety systems*	*Prep and disposal*	*Formal assessment*	*Abnormal events*	*Rules*	*Traction theory*	*Traction Practical*	
Conscientiousness	*r*	.05	.05	.12	.08		**.32**	**.49**		.20	-.08			**.69**	**.63**	**.59**	**.36**
	p	.38	.39	.22	.31		.02	<.001		.10	.32			<.001	<.001	<.01	<.01
	N	43	43	43	43		43	43		43	42			21	21	21	53
DCS	*r*									.10	-.05						**.30**
	p									.26	.38						.02
	N									43	42						53
TLS	*r*	.11	.11	.13	-.04	-.12	.12	**.56**	.05								
	p	.25	.25	.21	.41	.22	.23	<.001	.38								
	N	43	43	43	43	43	43	43	43								

Table 160 - SJE version A sub-criteria and manager ratings correlations table

SJE measure		Manager ratings of behaviour													
		Dependability	*Attitude to work and people*	*Commitment to work*	*Attention to detail*	*Checking and not making assumptions*	*Rule compliance*	*Proactivity*	*Tenacity*	*Calmness under pressure*	*Reactivity to stress*	*Assertiveness*	*Social need*	*Sensation seeking*	*Need for stimulation*
Dependability	*r*	**.24**													
	p	.03													
	N	59													
Attitude to work and people	*r*		**.23**												
	p		.04												
	N		59												
Commitment to work	*r*			**.23**											
	p			.04											
	N			59											
Attention to detail	*r*				**.28**										
	p				.02										

SJE measure		Manager ratings of behaviour													
		Dependability	*Attitude to work and people*	*Commitment to work*	*Attention to detail*	*Checking and not making assumptions*	*Rule compliance*	*Proactivity*	*Tenacity*	*Calmness under pressure*	*Reactivity to stress*	*Assertiveness*	*Social need*	*Sensation seeking*	*Need for stimulation*
	N				59										
Checking and not making assumptions	*r*					.02									
	p					.44									
	N					59									
Rule compliance	*r*						**.27**								
	p						.02								
	N						59								
Proactivity	*r*							**.38**							
	p							<.01							
	N							59							
Tenacity	*r*								**.26**						
	p								.03						
	N								59						
Calmness under pressure	*r*									**.24**					
	p									.04					
	N									59					
Reactivity to stress	*r*										**.25**				
	p										.03				
	N										59				
Assertiveness	*r*											**.37**			
	p											<.005			
	N											59			
Social need	*r*												**.26**		
	p												.03		
	N												59		
Sensation seeking	*r*													**.24**	
	p													.04	
	N													59	

SJE measure		Manager ratings of behaviour													
		Dependability	*Attitude to work and people*	*Commitment to work*	*Attention to detail*	*Checking and not making assumptions*	*Rule compliance*	*Proactivity*	*Tenacity*	*Calmness under pressure*	*Reactivity to stress*	*Assertiveness*	*Social need*	*Sensation seeking*	*Need for stimulation*
Need for stimulation	*r*														**.29**
	p														.01
	N														59

G.10 MMI

Table 161- Correlations between MMI main criteria scores and performance data

MMI Measure		Operational performance data												Training performance data		
		Train handling	*Rules*	*SPAD record*	*SPAD risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety systems*	*Prep and disposal*	*Formal assessment*	*Abnormal events*	*Rules*	*Traction theory*	*Traction Practical*
Conscientiousness	*r*	**.22**	.10	**.52**	**-.26**		.13	-.06		**.46**	-.19			.27	.19	.20
	p	.04	.22	<.01	.02		.15	.31		<.001	.06			.09	.16	.15
	N	68	68	68	68		68	68		68	68			28	28	28
DCS	*r*									**.29**	-.12					
	p									.01	.16					
	N									68	68					
TLS	*r*	.14	.15	.02	.04	.02	-.02	-.01	.12							
	p	.12	.11	.45	.36	.44	.44	.48	.16							
	N	68	68	68	68	68	68	68	68							

G.11 CBI

Table 162 - Correlations between CBI scores and performance data

CBI Measure		Operational performance data											Training performance data		
		Train handling	*Rules*	*SPAD record*	*SPAD Risk*	*Collisions*	*Speeding*	*Overruns*	*Disregards*	*Safety Systems*	*Prep and disposal*	*Formal assessment*	*Abnormal events*	*Rules*	*Traction theory*	*Traction practical*
Follows set rules and procedures	*r*	**-.25**	-.18	-.19	.19		-.06	-.06		.04	-.07			.19	-.17	.19
	P	.03	.08	.07	.07		.32	.33		.39	.30			.07	.21	.18
	N	62	62	62	62		62	62		62	62			62	25	25
Conscientiously works to meet training and job demands	*R*	-.13	-.03	-.06	.06		<.01	<.01		.16	.06			.06	-.01	.25
	p	.15	.42	.32	.32		.49	.49		.11	.33			.32	.49	.12
	N	62	62	62	62		62	62		62	62			62	25	25
Remain calm in emergency and stressful situations	*r*									.07			-.10			
	P									.30			.21			
	N									62			62			
Proactive and tenacious	*r*									.09			**-.37**			
	P									.29			.01			
	N									39			39			
Can spend time alone and does so effectively	*r*	**-**.22		-.21	.21	**-.29**		-.06	.12							
	p	.09		.10	.10	.04		.36	.24							
	N	39		39	39	39		39	39							

H. Inter-correlations between all assessment method scores

Table 163 – Inter-correlations of cognitive and psychomotor assessment methods

			2HAND				VIGIL			WAFV			TAVTMB		TEA-Occ		Group Bourdon			DTG			TRP	
			Total mean duration	*Total mean error duration*	*Total percent error duration*	*Coordination difficulty*	*Number of correct*	*Number of incorrect*	*Mean value of reaction time correct (sec.)*	*Number of 'missed reactions'*	*Mean reaction time*	*Number of 'false alarms'*	*Overview*	*Working time*	*Lift counting with distraction, total number of correctly counted strings*	*Tel search w. counting - dual task decrement*	*Production total*	*Omissions total*	*Faults total*	*Part 3 good*	*Part 3 wrong*	*Self paced wrong*	*Part 1*	*Part 2*
2HAND	*Total mean duration*	*r*		.059	-.402**	.427**	.117	-.095	-.123	-.085	-.063	.074	-.633*	-.162	-.706	-.171	-.012	.056	.137	.003	.073	-.044	-.070	-.064
		P		.515	.000	.000	.201	.299	.180	.352	.495	.419	.027	.615	.294	.829	.944	.749	.432	.982	.746	.832	.608	.640
		N		123	123	123	122	122	121	121	121	121	12	12	4	4	35	35	35	48	22	26	56	56
	Total mean error duration	*r*	.059		.787**	-.002	-.250**	-.041	-.021	.212*	.151	.071	.153	.407	-.853	-.394	-.288	-.074	-.213	-.237	-.135	.349	-.236	-.210
		P	.515		.000	.985	.005	.655	.823	.019	.099	.442	.635	.189	.147	.606	.094	.672	.220	.105	.548	.080	.080	.120
		N	123		123	123	122	122	121	121	121	121	12	12	4	4	35	35	35	48	22	26	56	56
	Total percent error duration	*r*	-.402**	.787**		-.192*	-.395**	-.033	.058	.374**	.271**	.002	.600*	.627*	-.280	-.325	-.274	-.079	-.183	-.132	.038	.348	-.164	-.142
		P	.000	.000		.033	.000	.720	.528	.000	.003	.979	.039	.029	.720	.675	.111	.653	.294	.371	.866	.081	.227	.297
		N	123	123		123	122	122	121	121	121	121	12	12	4	4	35	35	35	48	22	26	56	56
	2HAND Coordination difficulty	*r*	.427**	-.002	-.192*		-.016	-.093	-.040	-.053	-.061	.003	-.213	-.319	.540	.957*	-.027	-.154	.032	.146	-.131	.023	.117	.225
		P	.000	.985	.033		.863	.309	.663	.566	.507	.974	.506	.312	.460	.043	.877	.376	.854	.321	.560	.911	.390	.096
		N	123	123	123		122	122	121	121	121	121	12	12	4	4	35	35	35	48	22	26	56	56
VIGIL	*Number of correct*	*r*	.117	-.250**	-.395**	-.016		-.061	-.222*	-.250**	-.166	.020	-.612*	-.448	-.080	.255	.005	.015	.196	.137	-.218	-.054	.199	.294*
		P	.201	.005	.000	.863		.505	.014	.005	.067	.825	.034	.144	.920	.745	.976	.933	.268	.359	.343	.795	.145	.029
		N	122	122	122	122		123	122	122	122	122	12	12	4	4	34	34	34	47	21	26	55	55
	Number of incorrect	*r*	-.095	-.041	-.033	-.093	-.061		-.303**	-.024	-.011	.029	.051	.155	-.935	.136	-.133	-.075	-.020	-.021	-.134	.079	.025	.103
		P	.299	.655	.720	.309	.505		.001	.797	.905	.752	.875	.631	.065	.864	.452	.673	.910	.888	.561	.702	.856	.453
		N	122	122	122	122	123		122	122	122	122	12	12	4	4	34	34	34	47	21	26	55	55
	Mean value of reaction time correct (sec.)	*r*	-.123	-.021	.058	-.040	-.222*	-.303**		-.013	.316**	.008	.244	.376	.596	-.686	.301	-.168	-.140	.034	.396	.091	-.143	-.210
		P	.180	.823	.528	.663	.014	.001		.891	.000	.930	.470	.255	.593	.518	.089	.350	.436	.825	.084	.657	.301	.127
		N	121	121	121	121	122	122		121	121	121	11	11	3	3	33	33	33	46	20	26	54	54
WAFV	*Number of*	*r*	-.085	.212*	.374**	-.053	-.250**	-.024	-.013		.587**	.436**	.634*	.637*	-.149	-.401	-.286	.077	-.112	-.145	.266	-.291	-.102	-.054

			2HAND				VIGIL			WAFV			TAVTMB		TEA-Occ		Group Bourdon			DTG			TRP	
			Total mean duration	*Total mean error duration*	*Total percent error duration*	*Coordination difficulty*	*Number of correct*	*Number of incorrect*	*Mean value of reaction time correct (sec.)*	*Number of 'missed reactions'*	*Mean reaction time*	*Number of 'false alarms'*	*Overview*	*Working time*	*Lift counting with distraction, total number of correctly counted strings*	*Tel search w. counting - dual task decrement*	*Production total*	*Omissions total*	*Faults total*	*Part 3 good*	*Part 3 wrong*	*Self paced wrong*	*Part 1*	*Part 2*
	'missed reactions'	*P*	.352	.019	.000	.566	.005	.797	.891		.000	.000	.027	.026	.851	.599	.101	.665	.529	.332	.244	.149	.460	.695
		N	121	121	121	121	122	122	121		122	122	12	12	4	4	34	34	34	47	21	26	55	55
	Mean reaction time	*r*	-.063	.151	.271**	-.061	-.166	-.011	.316**	.587**		.292**	.577*	.808**	-.362	-.613	-.071	.165	-.047	-.145	.350	-.083	-.187	-.122
		P	.495	.099	.003	.507	.067	.905	.000	.000		.001	.049	.001	.638	.387	.692	.352	.790	.332	.120	.686	.172	.374
		N	121	121	121	121	122	122	121	122		122	12	12	4	4	34	34	34	47	21	26	55	55
	Number of "false alarms'	*r*	.074	.071	.002	.003	.020	.029	.008	.436**	.292**		-.413	-.002	-.135	-.805	-.185	.165	-.074	-.251	-.203	-.269	-.030	.054
		P	.419	.442	.979	.974	.825	.752	.930	.000	.001		.182	.995	.865	.195	.294	.350	.676	.089	.377	.184	.828	.697
		N	121	121	121	121	122	122	121	122	122		12	12	4	4	34	34	34	47	21	26	55	55
TAVTMB	*Overview*	*r*	-.633*	.153	.600*	-.213	-.612*	.051	.244	.634*	.577*	-.413		-.094	.256**	-.287**	-.102	-.079	-.188	.229*	.039	-.201	.231*	.202
		P	.027	.635	.039	.506	.034	.875	.470	.027	.049	.182		.162	.002	.001	.402	.515	.119	.035	.802	.207	.034	.063
		N	12	12	12	12	12	12	11	12	12	12		225	142	142	70	70	70	85	43	41	85	85
	Working time	*r*	-.162	.407	.627*	-.319	-.448	.155	.376	.637*	.808**	-.002	-.094		-.207*	.151	-.058	-.016	.153	-.146	.043	.161	-.012	-.070
		P	.615	.189	.029	.312	.144	.631	.255	.026	.001	.995	.162		.013	.074	.632	.895	.205	.182	.785	.314	.917	.523
		N	12	12	12	12	12	12	11	12	12	12	225		142	142	70	70	70	85	43	41	85	85
TEA-Occ	*Lift counting with distraction*	*r*	-.706	-.853	-.280	.540	-.080	-.935	.596	-.149	-.362	-.135	.256**	-.207*		-.392**	.098	-.141	-.086	.072	-.142	-.145	.094	.152
		P	.294	.147	.720	.460	.920	.065	.593	.851	.638	.865	.002	.013		.000	.388	.212	.449	.495	.541	.233	.372	.148
		N	4	4	4	4	4	4	3	4	4	4	142	142		177	80	80	80	91	21	69	93	92
	Tel search w. counting - dual task decrement	*r*	-.171	-.394	-.325	.957*	.255	.136	-.686	-.401	-.613	-.805	-.287**	.151	-.392**		-.028	.047	.040	-.330**	.264	.404**	-.143	-.337**
		P	.829	.606	.675	.043	.745	.864	.518	.599	.387	.195	.001	.074	.000		.807	.682	.726	.001	.247	.001	.170	.001
		N	4	4	4	4	4	4	3	4	4	4	142	142	177		80	80	80	91	21	69	93	92
Group Bourdon	*Production total*	*r*	-.012	-.288	-.274	-.027	.005	-.133	.301	-.286	-.071	-.185	-.102	-.058	.098	-.028		-.004	.003	.345**	-.037	.305**	.065	.041
		P	.944	.094	.111	.877	.976	.452	.089	.101	.692	.294	.402	.632	.388	.807		.965	.977	.000	.803	.007	.454	.636
		N	35	35	35	35	34	34	33	34	34	34	70	70	80	80		137	137	124	47	77	136	135

			2HAND				VIGIL			WAFV			TAVTMB		TEA-Occ		Group Bourdon			DTG			TRP	
			Total mean duration	*Total mean error duration*	*Total percent error duration*	*Coordination difficulty*	*Number of correct*	*Number of incorrect*	*Mean value of reaction time correct (sec.)*	*Number of 'missed reactions'*	*Mean reaction time*	*Number of 'false alarms'*	*Overview*	*Working time*	*Lift counting with distraction, total number of correctly counted strings*	*Tel search w. counting - dual task decrement*	*Production total*	*Omissions total*	*Faults total*	*Part 3 good*	*Part 3 wrong*	*Self paced wrong*	*Part 1*	*Part 2*
	Omissions total	*r*	.056	-.074	-.079	-.154	.015	-.075	-.168	.077	.165	.165	-.079	-.016	-.141	.047	-.004		-.049	.152	.111	.004	-.168	-.065
		P	.749	.672	.653	.376	.933	.673	.350	.665	.352	.350	.515	.895	.212	.682	.965		.572	.092	.456	.969	.051	.455
		N	35	35	35	35	34	34	33	34	34	34	70	70	80	80	137		137	124	47	77	136	135
	Faults total	*r*	.137	-.213	-.183	.032	.196	-.020	-.140	-.112	-.047	-.074	-.188	.153	-.086	.040	.003	-.049		-.036	-.153	.097	-.138	-.099
		P	.432	.220	.294	.854	.268	.910	.436	.529	.790	.676	.119	.205	.449	.726	.977	.572		.691	.304	.400	.110	.252
		N	35	35	35	35	34	34	33	34	34	34	70	70	80	80	137	137		124	47	77	136	135
DTG	*Part 3 good*	*r*	.003	-.237	-.132	.146	.137	-.021	.034	-.145	-.145	-.251	.229*	-.146	.072	-.330**	.345**	.152	-.036		-.371**	-.173	.093	.214**
		P	.982	.105	.371	.321	.359	.888	.825	.332	.332	.089	.035	.182	.495	.001	.000	.092	.691		.002	.093	.236	.006
		N	48	48	48	48	47	47	46	47	47	47	85	85	91	91	124	124	124		68	95	163	162
	Part 3 wrong	*r*	.073	-.14	.04	-.131	-.218	-.134	.396	.266	.350	-.203	.039	.043	-.142	.264	-.037	.111	-.153	-.371**		.c	-.023	-.043
		P	.746	.548	.866	.560	.343	.561	.084	.244	.120	.377	.802	.785	.541	.247	.803	.456	.304	.002		.	.854	.726
		N	22	22	22	22	21	21	20	21	21	21	43	43	21	21	47	47	47	68		0	68	68
	Self-paced wrong	*R*	-.044	.349	.348	.023	-.054	.079	.091	-.291	-.083	-.269	-.201	.161	-.145	.404**	.305**	.004	.097	-.173	.c		-.157	-.229*
		P	.832	.080	.081	.911	.795	.702	.657	.149	.686	.184	.207	.314	.233	.001	.007	.969	.400	.093	.		.130	.028
		N	26	26	26	26	26	26	26	26	26	26	41	41	69	69	77	77	77	95	0		94	93
TRP	*Part 1*	*r*	-.070	-.236	-.164	.117	.199	.025	-.143	-.102	-.187	-.030	.231*	-.012	.094	-.143	.065	-.168	-.138	.093	-.023	-.157		.655**
		P	.608	.080	.227	.390	.145	.856	.301	.460	.172	.828	.034	.917	.372	.170	.454	.051	.110	.236	.854	.130		.000
		N	56	56	56	56	55	55	54	55	55	55	85	85	93	93	136	136	136	163	68	94		174
	Part 2	*r*	-.064	-.210	-.142	.225	.294*	.103	-.210	-.054	-.122	.054	.202	-.070	.152	-.337**	.041	-.065	-.099	.214**	-.043	-.229*	.655**	
		P	.640	.120	.297	.096	.029	.453	.127	.695	.374	.697	.063	.523	.148	.001	.636	.455	.252	.006	.726	.028	.000	
		N	56	56	56	56	55	55	54	55	55	55	85	85	92	92	135	135	135	162	68	93	174	

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

c. Cannot be computed because at least one of the variables is constant.

Table 164 – Pearson correlations between SJE, MMI and CBI and all other recommended assessment measures

			SJE			MMI				CBI					
			Consc.	*DCS*	*TLS*	*Consc.*	*DCS*	*TLS*	*Verbal comm.*	*Rules and procedures*	*Consc.*	*Proactive and tenacious*	*Calm in emergency and stressful situations*	*Can spend time alone and does so effectively*	*Comm.*
2HAND	*Total mean duration*	*r*	-.02	.02	-.06	.04	.17	-.08	.01	-.13	-.08	.19	**-.32**	-.06	-.13
		P	.88	.88	.63	.75	.16	.50	.96	.37	.58	.26	.03	.74	.39
		N	57	57	57	75	75	75	74	48	48	37	48	37	48
	Total mean error duration	*r*	-.20	-.10	**-.30**	-.09	-.05	-.04	-.04	.17	-.10	-.28	.08	-.21	.11
		P	.13	.48	.02	.44	.67	.75	.75	.25	.52	.09	.61	.21	.44
		N	57	57	57	75	75	75	74	48	48	37	48	37	48
	Total percent error duration	*r*	-.03	-.01	-.15	-.08	-.12	.02	-.04	.19	-.06	-.24	.09	-.16	.16
		P	.84	.95	.26	.50	.31	.86	.74	.20	.67	.15	.56	.36	.27
		N	57	57	57	75	75	75	74	48	48	37	48	37	48
	2HAND Coordination difficulty	*r*	-.11	-.05	-.24	-.02	-.03	-.09	**-.29**	-.13	-.13	.13	-.12	-.11	-.16
		P	.43	.69	.07	.87	.77	.43	.01	.38	.39	.45	.43	.54	.28
		N	57	57	57	75	75	75	74	48	48	37	48	37	48
VIGIL	*Number of correct*	*r*	.02	.08	.04	.06	.12	.06	.00	.07	.09	-.03	.16	.14	.10
		P	.91	.54	.77	.61	.29	.59	.99	.63	.53	.85	.28	.40	.50
		N	56	56	56	76	76	76	75	47	47	36	47	36	47
	Number of incorrect	*r*	-.22	-.14	-.09	-.06	-.05	.10	.07	.13	.18	-.10	-.14	-.13	.15
		P	.10	.31	.54	.58	.66	.37	.53	.38	.22	.57	.36	.46	.31
		N	56	56	56	76	76	76	75	47	47	36	47	36	47
	Mean value of reaction time correct (sec.)	*r*	.19	.12	.12	.16	.15	.02	-.04	-.07	-.10	.25	.17	-.04	-.07
		P	.16	.38	.40	.18	.21	.83	.77	.65	.51	.15	.25	.80	.63
		N	55	55	55	76	76	76	75	46	46	35	46	35	46
WAFV	*Number of 'missed reactions'*	*r*	-.03	-.03	-.06	.00	.02	.05	.09	.12	-.07	-.09	-.10	-.18	.08
		P	.81	.85	.64	.99	.87	.66	.43	.43	.63	.62	.50	.30	.62
		N	56	56	56	75	75	75	74	47	47	36	47	36	47
	Mean reaction time	*r*	-.06	-.07	-.06	-.02	.10	.03	.08	.05	.10	.09	-.15	-.15	.18

			SJE			MMI				CBI					
			Consc.	*DCS*	*TLS*	*Consc.*	*DCS*	*TLS*	*Verbal comm.*	*Rules and procedures*	*Consc.*	*Proactive and tenacious*	*Calm in emergency and stressful situations*	*Can spend time alone and does so effectively*	*Comm.*
		P	.66	.60	.68	.90	.40	.81	.52	.75	.50	.58	.31	.37	.22
		N	56	56	56	75	75	75	74	47	47	36	47	36	47
	Number of 'false alarms'	*r*	-.19	-.09	-.07	-.01	.03	.12	.03	-.16	-.11	-.12	-.10	-.18	-.17
		P	.17	.50	.60	.92	.78	.30	.79	.28	.48	.50	.48	.29	.26
		N	56	56	56	75	75	75	74	47	47	36	47	36	47
TAVTMB	*Overview*	*r*	.98	.77	-.03	.06	-.28	-.55	-.13	.19	-.02	.10	.13	.07	**.26**
		P	.13	.44	.98	.90	.51	.16	.75	.09	.86	.42	.23	.57	.02
		N	3	3	3	8	8	8	8	82	82	67	82	67	82
	Working time	*r*	.98	.77	-.02	.69	.40	.32	-.16	.09	.06	-.05	.04	.18	.09
		P	.13	.44	.99	.06	.32	.44	.70	.40	.60	.67	.71	.15	.42
		N	3	3	3	8	8	8	8	82	82	67	82	67	82
TEA-Occ	*Lift counting with distraction*	*r*	**-1.00**	**1.00**[*]	**1.00**	**-.10**[*]	-.92	-.37	.[c]	.09	-.07	.15	.09	.04	.11
		P	<.001	<.001	<.001	.04	.25	.76	.00	.39	.53	.17	.42	.72	.32
		N	2	2	2	3	3	3	3	88	88	87	88	87	88
	Tel search w. counting - dual task decrement	*r*	**1.00**	**-1.00**	**-1.00**	-.24	-.54	-.98	.[c]	-.04	.11	-.01	**.22**	.21	-.15
		P	<.001	<.001	<.001	.85	.64	.13	.00	.69	.32	.91	.04	.05	.16
		N	2	2	2	3	3	3	3	88	88	87	88	87	88
Group Bourdon	*Production total*	*r*	-.12	-.18	.25	.06	-.25	.29	.01	.14	**.18**	.07	.15	.08	.16
		P	.61	.47	.30	.80	.27	.21	.96	.13	<.05	.47	.09	.42	.07
		N	19	19	19	21	21	21	21	124	124	101	124	101	124
	Omissions total	*r*	.22	-.06	.33	-.34	-.33	-.02	.03	-.17	.00	.03	-.02	.02	-.11
		P	.36	.80	.17	.13	.14	.93	.91	.06	1.00	.76	.79	.85	.24
		N	19	19	19	21	21	21	21	124	124	101	124	101	124
	Faults total	*r*	.05	.06	.19	-.35	-.06	-.36	.09	-.12	-.09	-.10	**-.18**	-.03	-.12
		P	.84	.81	.44	.12	.80	.11	.69	.19	.32	.34	.05	.74	.17
		N	19	19	19	21	21	21	21	124	124	101	124	101	124
D	*Part 3 good*	*r*	.09	.07	.07	.08	-.04	.23	.13	.04	-.07	.05	-.05	-.02	.03

			SJE			MMI				CBI					
			Consc.	*DCS*	*TLS*	*Consc.*	*DCS*	*TLS*	*Verbal comm.*	*Rules and procedures*	*Consc.*	*Proactive and tenacious*	*Calm in emergency and stressful situations*	*Can spend time alone and does so effectively*	*Comm.*
		P	.63	.71	.74	.64	.79	.16	.42	.62	.37	.58	.55	.85	.72
		N	28	28	28	39	39	39	38	161	161	133	161	133	161
	Part 3 wrong	*r*	.05	.04	-.17	-.34	**-.50**	-.35	-.35	.07	-.01	.05	.17	.26	-.04
		P	.88	.90	.61	.14	.02	.12	.12	.58	.92	.77	.17	.11	.77
		N	12	12	12	21	21	21	21	67	67	39	67	39	67
	Self-paced wrong	*R*	-.47	-.52*	-.45	.42	.20	.25	-.10	.13	-.18	-.13	.08	.17	-.06
		P	.07	.04	.08	.08	.42	.31	.71	.20	.08	.21	.46	.11	.60
		N	16	16	16	18	18	18	17	93	93	93	93	93	93
TRP	*Part 1*	*r*	.04	.13	-.05	-.07	-.21	-.05	.16	.20*	-.01	.03	.11	.05	.12
		P	.81	.49	.79	.65	.19	.74	.33	.01	.94	.70	.18	.54	.13
		N	33	33	33	41	41	41	40	160	160	132	160	132	160
	Part 2	*r*	-.03	-.07	-.15	.04	-.11	-.03	.09	.08	-.08	.05	-.02	-.09	.13
		P	.86	.69	.41	.78	.50	.86	.59	.33	.29	.59	.83	.29	.09
		N	33	33	33	41	41	41	40	159	159	131	159	131	159
SJE	*Conscientiousness*	*R*		**.85**	**.78**	.30*	.25	.20	-.07	.12	.06	.16	.24	.23	-.10
		P		<.001	<.001	.05	.10	.19	.65	.55	.76	.45	.22	.29	.62
		N		69	69	44	44	44	44	28	28	24	28	24	28
	DCS	*R*	**.85**		**.69**	.11	.08	.06	-.03	.32	.13	.04	.35	.34	-.03
		P	.00		.00	.47	.61	.70	.87	.10	.52	.84	.07	.11	.89
		N	69		69	44	44	44	44	28	28	24	28	24	28
	TLS	*R*	**.78**	**.69**		.20	.14	-.04	.09	.15	.12	-.07	.13	.36	-.09
		P	<.001	<.001		.19	.38	.79	.57	.45	.56	.76	.50	.09	.66
		N	69	69		44	44	44	44	28	28	24	28	24	28
MMI	*Conscientiousness*	*R*	**.30**	.11	.20		**.76**	**.36**	-.07	.15	.10	-.01	.20	.16	.14
		P	.05	.47	.19		.00	.00	.50	.38	.57	.95	.23	.40	.41
		N	44	44	44		92	92	91	39	39	30	39	30	39
	DCS	*R*	.25	.08	.14	**.76***		**.46**	-.07	.30	.17	.14	.12	.23	.18

			SJE			MMI				CBI					
			Consc.	*DCS*	*TLS*	*Consc.*	*DCS*	*TLS*	*Verbal comm.*	*Rules and procedures*	*Consc.*	*Proactive and tenacious*	*Calm in emergency and stressful situations*	*Can spend time alone and does so effectively*	*Comm.*
		P	.10	.61	.38	<.001		<.001	.49	.06	.31	.45	.47	.23	.26
		N	44	44	44	92		92	91	39	39	30	39	30	39
	TLS	*R*	.20	.06	-.04	**.36**	**.46**		-.14	.11	**.33**	.11	.02	.22	.21
		P	.19	.70	.79	<.001	<.001		.20	.49	.04	.56	.91	.24	.19
		N	44	44	44	92	92		91	39	39	30	39	30	39
	Verbal communication	*r*	-.07	-.03	.09	-.07	-.07	-.14		-.14	-.07	-.28	-.29	-.15	.28
		P	.65	.87	.57	.50	.49	.20		.41	.68	.14	.08	.45	.09
		N	44	44	44	91	91	91		38	38	29	38	29	38
CBI	*Follows set rules and procedures*	*r*	.12	.32	.15	.15	.30	.11	-.14		**.38**	**.24**	**.52**	**.56**	**.32**
		P	.55	.10	.45	.38	.06	.49	.41		<.001	.01	<.001	<.001	<.001
		N	28	28	28	39	39	39	38		161	133	161	133	161
	Conscientiousness	*r*	.06	.13	.12	.10	.17	**.33**	-.07	**.38**		**.50**	**.31**	**.40**	.31
		P	.76	.52	.56	.57	.31	.04	.68	<.001		<.001	<.001	<.001	<.001
		N	28	28	28	39	39	39	38	161		133	161	133	161
	Proactive and tenacious	*r*	.16	.04	-.07	-.01	.14	.11	-.28	**.24**	**.50**		**.38**	**.41**	**.20**
		P	.45	.84	.76	.95	.45	.56	.14	.01	<.001		<.001	<.001	.02
		N	24	24	24	30	30	30	29	133	133		133	133	133
	Calm in emergency	*r*	.24	.35	.13	.20	.12	.02	-.29	**.52**	**.31**	**.38**		**.53**	**.26**
		P	.22	.07	.50	.23	.47	.91	.08	<.001	<.001	<.001		<.001	<.001
		N	28	28	28	39	39	39	38	161	161	133		133	161
	Spend time alone	*r*	.23	.34	.36	.16	.23	.22	-.15	**.56**	**.40**	**.41**	**.53**		**.23**
		P	.29	.11	.09	.40	.23	.24	.45	<.001	<.001	<.001	<.001		.01
		N	24	24	24	30	30	30	29	133	133	133	133		133
	Communication	*r*	-.10	-.03	-.09	.14	.18	.21	.28	**.32**	**.31**	**.20**	**.25**	**.23**	
		P	.62	.89	.66	.41	.26	.19	.09	<.001	<.001	.02	<.001	.01	
		N	28	28	28	39	39	39	38	161	161	133	161	133	
WC	*Accuracy*	*r*	.32	.47	.11	.04	-.03	-.13	.06	-.10	-.28	-.39	-.16	-.31	.13

			SJE			MMI				CBI					
			Consc.	*DCS*	*TLS*	*Consc.*	*DCS*	*TLS*	*Verbal comm.*	*Rules and procedures*	*Consc.*	*Proactive and tenacious*	*Calm in emergency and stressful situations*	*Can spend time alone and does so effectively*	*Comm.*
		P	.08	.01	.54	.82	.84	.40	.71	.62	.15	.11	.42	.21	.52
		N	31	31	31	45	45	45	44	28	28	18	28	18	28
	Written comprehension	*r*	.37*	.30	.20	.22	.09	.02	-.02	.14	.16	.12	.05	.04	.26
		P	.04	.11	.29	.15	.54	.92	.91	.47	.41	.63	.81	.87	.18
		N	31	31	31	45	45	45	44	28	28	18	28	18	28
	Legibility	*r*	>-.001	.01	-.06	-.09	-.01	-.08	.17	-.21	-.26	-.30	-.26	-.24	.06
		P	.99	.97	.73	.55	.95	.61	.28	.29	.19	.22	.19	.33	.76
		N	31	31	31	45	45	45	44	28	28	18	28	18	28
	Structure	*r*	.12	.22	-.09	-.07	.06	.05	-.10	.08	.30	.19	.02	.24	-.22
		P	.51	.23	.64	.66	.71	.77	.51	.68	.12	.44	.92	.33	.25
		N	31	31	31	45	45	45	44	28	28	18	28	18	28
	Overall	*r*	.51**	.54**	.23	.18	.07	-.06	.02	.08	.05	-.10	-.04	-.11	.26
		P	>.001	>.001	.22	.23	.64	.72	.88	.70	.79	.71	.84	.67	.18
		N	31	31	31	45	45	45	44	28	28	18	28	18	28
WCT version 2	*Accuracy*	*r*	.c	.c	.c	-.07	.09	.15	.10	-.16	.11	.13	-.20	-.24	.22
		P	<.001	<.001	<.001	.65	.56	.32	.52	.44	.58	.57	.33	.26	.30
		N	37	37	37	45	45	45	45	26	26	23	26	23	26
	Written comprehension	*r*	.31	.18	.27	.26	.08	-.12	.13	-.28	-.34	-.39	-.37	-.28	-.34
		P	.06	.28	.11	.09	.60	.45	.38	.16	.09	.07	.06	.19	.09
		N	37	37	37	45	45	45	45	26	26	23	26	23	26
	Legibility	*r*	.31	.49**	.50**	-.05	-.05	-.16	.39**	.37	.17	.18	.30	.27	.02
		P	.06	<.001	<.001	.75	.74	.30	.01	.07	.42	.40	.15	.21	.91
		N	37	37	37	45	45	45	45	26	26	23	26	23	26
	Structure	*r*	.12	.19	.10	.03	-.13	.01	.35*	.02	-.27	.01	.10	.40	-.06
		P	.49	.25	.58	.86	.40	.96	.02	.94	.19	.96	.64	.06	.77
		N	37	37	37	45	45	45	45	26	26	23	26	23	26
	Overall	*r*	.39*	.31	.37*	.22	.05	-.11	.29	-.21	-.35	-.30	-.30	-.14	-.30

			SJE			MMI				CBI					
			Consc.	*DCS*	*TLS*	*Consc.*	*DCS*	*TLS*	*Verbal comm.*	*Rules and procedures*	*Consc.*	*Proactive and tenacious*	*Calm in emergency and stressful situations*	*Can spend time alone and does so effectively*	*Comm.*
		P	.02	.06	.03	.15	.76	.45	.06	.30	.08	.16	.14	.54	.13
		N	37	37	37	45	45	45	45	26	26	23	26	23	26

Table 165 – Pearson correlations between WCT version 1 and 2 and all other recommended assessment measures

			WCT version 1					WCT version 2				
			Accuracy	*Written comprehension*	*Legibility*	*Structure*	*Overall*	*Accuracy*	*Written comprehension*	*Legibility*	*Structure*	*Overall*
2HAND	*Total mean duration*	*r*	-.057	-.061	.096	.190	-.041	.252*	.158	-.069	.198	.201
		P	.667	.645	.465	.145	.758	.050	.225	.595	.126	.120
		N	60	60	60	60	60	61	61	61	61	61
	Total mean error duration	*r*	-.090	-.052	-.005	.005	-.072	-.171	-.095	-.123	.162	-.100
		P	.493	.694	.967	.969	.584	.188	.464	.346	.211	.441
		N	60	60	60	60	60	61	61	61	61	61
	Total percent error duration	*r*	-.094	.038	-.027	-.088	-.004	-.356**	-.077	.003	.090	-.095
		P	.476	.772	.840	.505	.974	.005	.556	.982	.488	.468
		N	60	60	60	60	60	61	61	61	61	61
	2HAND Coordination difficulty	*r*	.104	-.034	.083	.082	.006	-.090	-.065	-.205	.078	-.091
		P	.431	.798	.529	.534	.965	.489	.620	.113	.553	.484
		N	60	60	60	60	60	61	61	61	61	61
VIGIL	*Number of correct*	*r*	.172	-.026	.114	-.002	.056	.284*	.139	.213	.066	.211
		P	.189	.844	.388	.990	.673	.026	.284	.100	.614	.103
		N	60	60	60	60	60	61	61	61	61	61
	Number of incorrect	*r*	-.051	-.124	.054	.070	-.116	.036	.077	.037	-.325*	.025
		P	.696	.345	.684	.593	.376	.780	.553	.777	.011	.847
		N	60	60	60	60	60	61	61	61	61	61
	Mean value of reaction time correct (sec.)	*r*	-.151	.067	-.032	-.144	-.010	.029	-.013	-.102	.024	-.023
		P	.248	.609	.806	.272	.938	.828	.919	.436	.858	.864
		N	60	60	60	60	60	60	60	60	60	60
WAFV	*Number of 'missed reactions'*	*r*	-.071	-.205	.099	.058	-.195	.043	-.149	.065	.047	-.120
		P	.589	.117	.450	.659	.135	.743	.256	.620	.724	.360
		N	60	60	60	60	60	60	60	60	60	60
	Mean reaction time	*r*	.028	.143	-.063	.050	.155	.126	-.154	-.032	.039	-.136
		P	.834	.275	.633	.705	.236	.338	.242	.806	.769	.300
		N	60	60	60	60	60	60	60	60	60	60

			WCT version 1					WCT version 2				
			Accuracy	*Written comprehension*	*Legibility*	*Structure*	*Overall*	*Accuracy*	*Written comprehension*	*Legibility*	*Structure*	*Overall*
	Number of 'false alarms'	*r*	.048	-.101	.070	.060	-.065	.053	-.180	-.029	.058	-.163
		P	.717	.441	.597	.651	.620	.685	.168	.826	.659	.213
		N	60	60	60	60	60	60	60	60	60	60
TAVTMB	*Overview*	*r*	.[b]	.556	.[b]	.[b]	.556	-.186	-.169	.[c]	.[c]	-.204
		P	0.000	.444	0.000	0.000	.444	.607	.641	0.000	0.000	.572
		N	4	4	4	4	4	10	10	10	10	10
	Working time	*r*	.[b]	-.318	.[b]	.[b]	-.318	.389	-.115	.[c]	.[c]	0.000
		P	0.000	.682	0.000	0.000	.682	.266	.753	0.000	0.000	1.000
		N	4	4	4	4	4	10	10	10	10	10
TEA-Occ	*Lift counting with distraction*	*r*	.[b]	.[b]	.[b]	.[b]	.[b]	-.293	-.535	.[c]	.[c]	-.608
		P						.573	.275	0.000	0.000	.201
		N	0	0	0	0	0	6	6	6	6	6
	Tel search w. counting - dual task decrement	*r*	.[b]	.[b]	.[b]	.[b]	.[b]	-.952**	.188	.[c]	.[c]	-.224
		P						.003	.721	0.000	0.000	.670
		N	0	0	0	0	0	6	6	6	6	6
Group Bourdon	*Production total*	*r*	.090	.152	-.134	-.345	.089	.317	.213	.[c]	-.001	.252
		P	.699	.512	.562	.125	.702	.186	.380	0.000	.995	.298
		N	21	21	21	21	21	19	19	19	19	19
	Omissions total	*r*	.120	.017	-.053	.003	.032	.161	.087	.[c]	.190	.155
		P	.605	.942	.819	.991	.889	.511	.723	0.000	.435	.525
		N	21	21	21	21	21	19	19	19	19	19
	Faults total	*r*	.085	-.273	-.362	.124	-.220	.081	.110	.[c]	-.322	.037
		P	.713	.231	.106	.593	.339	.742	.654	0.000	.179	.880
		N	21	21	21	21	21	19	19	19	19	19
DTG	*Part 3 good*	*r*	.125	.096	-.209	-.201	.087	-.015	.160	-.132	.180	.173
		P	.528	.626	.285	.305	.659	.942	.436	.521	.378	.398
		N	28	28	28	28	28	26	26	26	26	26
	Part 3 wrong	*r*	-.093	-.152	.236	.056	-.135	.[c]	-.158	.[c]	-.510	-.259
		P	.731	.574	.379	.836	.619	0.000	.642	0.000	.109	.442
		N	16	16	16	16	16	11	11	11	11	11
	Self-paced wrong	*R*	-.377	.407	-.207	-.434	.031	.050	-.017	-.244	.356	.053
		P	.227	.189	.518	.159	.923	.860	.952	.380	.192	.850
		N	12	12	12	12	12	15	15	15	15	15
TRP	*Part 1*	*r*	-.051	-.312	.157	-.441*	-.357*	-.167	.332	.023	.098	.330
		P	.780	.077	.384	.010	.041	.378	.073	.902	.607	.075
		N	33	33	33	33	33	30	30	30	30	30
	Part 2	*r*	.091	-.334	.133	-.119	-.282	.053	.210	.093	.046	.236

			WCT version 1					WCT version 2				
			Accuracy	*Written comprehension*	*Legibility*	*Structure*	*Overall*	*Accuracy*	*Written comprehension*	*Legibility*	*Structure*	*Overall*
		P	.613	.058	.461	.508	.111	.782	.266	.625	.809	.209
		N	33	33	33	33	33	30	30	30	30	30
SJE	*Conscientiousness*	*R*	.321	.369*	-.002	.124	.511**	.[c]	.314	.310	.118	.390*
		P	.078	.041	.992	.505	.003	0.000	.058	.062	.485	.017
		N	31	31	31	31	31	37	37	37	37	37
	DCS	*R*	.470**	.296	.008	.221	.540**	.[c]	.184	.494**	.193	.310
		P	.008	.106	.966	.233	.002	0.000	.276	.002	.253	.062
		N	31	31	31	31	31	37	37	37	37	37
	TLS	*R*	.114	.198	-.064	-.087	.225	.[c]	.265	.499**	.095	.368*
		P	.540	.285	.732	.641	.223	0.000	.113	.002	.576	.025
		N	31	31	31	31	31	37	37	37	37	37
MMI	*Conscientiousness*	*R*	.035	.218	-.091	-.066	.183	-.071	.255	-.049	.026	.217
		P	.819	.150	.551	.664	.228	.645	.091	.749	.864	.151
		N	45	45	45	45	45	45	45	45	45	45
	DCS	*R*	-.031	.093	-.010	.057	.071	.089	.081	-.050	-.128	.047
		P	.838	.543	.950	.709	.644	.559	.598	.742	.401	.758
		N	45	45	45	45	45	45	45	45	45	45
	TLS	*R*	-.130	.016	-.077	.045	-.055	.151	-.115	-.159	.007	-.114
		P	.395	.915	.613	.771	.720	.323	.452	.296	.964	.454
		N	45	45	45	45	45	45	45	45	45	45
	Verbal communication	*r*	.057	-.017	.168	-.102	.023	.098	.133	.393**	.350*	.285
		P	.714	.914	.277	.509	.884	.522	.384	.007	.018	.058
		N	44	44	44	44	44	45	45	45	45	45
CBI	*Follows set rules and procedures*	*r*	-.097	.142	-.207	.083	.077	-.158	-.282	.365	.015	-.214
		P	.622	.471	.291	.676	.698	.440	.162	.067	.940	.295
		N	28	28	28	28	28	26	26	26	26	26
	Conscientiousness	*r*	-.277	.161	-.258	.298	.054	.114	-.342	.165	-.265	-.347
		P	.154	.414	.185	.124	.787	.578	.088	.421	.191	.082
		N	28	28	28	28	28	26	26	26	26	26
	Proactive and tenacious	*r*	-.392	.121	-.304	.193	-.095	.127	-.386	.183	.011	-.302
		P	.108	.633	.220	.442	.708	.565	.069	.402	.959	.162
		N	18	18	18	18	18	23	23	23	23	23
	Calm in emergency	*r*	-.158	.048	-.258	.020	-.040	-.200	-.374	.289	.098	-.297
		P	.423	.808	.185	.920	.839	.327	.059	.153	.635	.140
		N	28	28	28	28	28	26	26	26	26	26
	Spend time alone	*r*	-.312	.042	-.243	.243	-.109	-.243	-.281	.271	.402	-.135
		P	.207	.867	.332	.332	.668	.264	.193	.212	.057	.539

			WCT version 1					WCT version 2				
			Accuracy	*Written comprehension*	*Legibility*	*Structure*	*Overall*	*Accuracy*	*Written comprehension*	*Legibility*	*Structure*	*Overall*
	Communication	*N*	18	18	18	18	18	23	23	23	23	23
		r	.126	.262	.061	-.223	.260	.216	-.342	.022	-.060	-.304
		P	.523	.177	.758	.254	.182	.289	.087	.914	.770	.130
		N	28	28	28	28	28	26	26	26	26	26

I. Pass rates for assessment scores if the recommended scoring rules are applied

Table 166 – Pass rates for each assessment score according to protected characteristics if recommended scoring rules were applied to trial sample[3]

Test	Score	Outcome	Protected characteristic																	
			White		Other		All ethnic group		Male		Female		All gender		50 and under		51+		All age	
			n	*%*	*n*	*%*	*Total n*	*4/5ths rule*	*n*	*%*	*n*	*%*	*Total n*	*4/5ths rule*	*n*	*%*	*n*	*%*	*Total n*	*4/5ths rule*
TEA-Occ	Lift counting with distraction	Pass	161	94%	4	67%	165	FAIL	160	93%	5	100%	165	OK	149	96%	16	76%	165	FAIL
		Fail	10	6%	2	33%	12		12	7%	0	0%	12		7	4%	5	24%	12	
		Total	171		6		177		172		5		177		156		21		177	
	Dual task decrement	Pass	157	92%	4	67%	161	FAIL	156	91%	5	100%	161	OK	144	92%	17	81%	161	OK
		Fail	14	8%	2	33%	16		16	9%	0	0%	16		12	8%	4	19%	16	
		Total	171		6		177		172		5		177		156		21		177	
Group Bourdon	Total production	Pass	1157	89%	333	90%	1490	OK	1389	89%	106	91%	1495	OK	1421	90%	65	76%	1486	OK
		Fail	145	11%	38	10%	183		173	11%	10	9%	183		160	10%	21	24%	181	
		Total	1302		371		1673		1562		116		1678		1581		86		1667	
	Total omissions	Pass	1216	93%	307	83%	1523	OK	1424	92%	104	90%	1528	OK	1438	91%	80	93%	1518	OK
		Fail	85	7%	64	17%	149		127	8%	12	10%	139		142	9%	6	7%	148	
		Total	1301		371		1672		1551		116		1667		1580		86		1666	
TRP1		Pass	1189	97%	213	85%	1402	OK	1317	94%	89	97%	1406	OK	1321	95%	75	99%	1396	OK
		Fail	41	3%	39	15%	80		77	6%	3	3%	80		75	5%	1	1%	76	
		Total	1230		252		1482		1394		92		1486		1396		76		1472	
TRP2		Pass	1127	91%	194	77%	1321	OK	1244	89%	81	88%	1325	OK	1249	89%	63	83%	1312	OK
		Fail	106	9%	57	23%	163		152	11%	11	12%	163		149	11%	13	17%	162	
		Total	1233		251		1484		1396		92		1488		1398		76		1474	
WAFV	Missed reactions	Pass	110	97%	9	100%	119	OK	108	97%	11	100%	119	OK	105	98%	14	93%	119	OK
		Fail	3	3%	0	0%	3		3	3%	0	0%	3		2	2%	1	7%	3	
		Total	113		9		122		111		11		122		107		15		122	
	False alarms	Pass	105	93%	6	67%	111	FAIL	101	91%	10	91%	111	OK	98	92%	13	87%	111	OK
		Fail	8	7%	3	33%	11		10	9%	1	9%	11		9	8%	2	13%	11	

[3] The statistics shown for the Group Bourdon, TRP 1 and TRP 2 were calculated by applying the recommended scoring rules to previous assessment data from financial year 2010/2011 because this provided a larger sample which consists of the genuine candidate population.

Test	Score	Outcome	Protected characteristic																	
			White		Other		All ethnic group		Male		Female		All gender		50 and under		51+		All age	
			n	*%*	*n*	*%*	*Total n*	*4/5ths rule*	*n*	*%*	*n*	*%*	*Total n*	*4/5ths rule*	*n*	*%*	*n*	*%*	*Total n*	*4/5ths rule*
		Total	113		9		122		111		11		122		107		15		122	
	Reaction time	Pass	112	99%	9	100%	121	OK	110	99%	11	100%	121	OK	106	99%	15	100%	121	OK
		Fail	1	1%	0	0%	1		1	1%	0	0%	1		1	1%	0	0%	1	
		Total	113		9		122		111		11		122		107		15		122	
TAVTMB	Overview	Pass	205	99%	10	100%	215	OK	216	100%	7	78%	223	FAIL	192	99%	30	100%	222	OK
		Fail	2	1%	0	0%	2		0	0%	2	22%	2		2	1%	0	0%	2	
		Total	207		10		217		216		9		225		194		30		224	
2HAND	Overall mean duration	Pass	113	99%	9	100%	122	OK	111	99%	11	100%	122	OK	107	99%	15	100%	122	OK
		Fail	1	1%	0	0%	1		1	1%	0	0%	1		1	1%	0	0%	1	
		Total	114		9		123		112		11		123		108		15		123	
	Percent error duration	Pass	111	97%	9	100%	120	OK	109	97%	11	100%	120	OK	105	97%	15	100%	120	OK
		Fail	3	3%	0	0%	3		3	3%	0	0%	3		3	3%	0	0%	3	
		Total	114		9		123		112		11		123		108		15		123	

Annex 4 - Independent review of the RSSB train driver selection research (Independent reviewers)

Winckworth Sherwood

Review of report on T948 Project on Driver Selection

Summarised answers to legal question based on extracts from legal advice letter

Background

Winckworth Sherwood Solicitors and Parliamentary Agents were appointed by RSSB to review the final report of research project T948 *Train Driver Selection* and the associated industry strategy for driver psychometric assessment from a legal perspective. The review had a particular focus on the impact of the recommended new psychometric assessment process on candidates in ethnic minority groups.

The relevant law relates to direct and indirect discrimination as it would apply to an employer of train drivers using psychometric assessment.

Direct discrimination occurs where an employer treats an employee or a job applicant less favourably because of a protected characteristic under the Equality Act 2010.

Indirect discrimination is concerned with acts, decisions or policies which are not intended to treat anyone less favourably, but which in practice have the effect of disadvantaging a group of people with a particular protected characteristic. Where such a policy disadvantages an individual with that characteristic, it will amount to indirect discrimination unless it can be objectively justified.

The statutory definition is found in section 19 of the Equality Act 2010. When applied to the psychometric assessment process, the employer would be considered to discriminate indirectly against a candidate where:

1. Employer applies a provision, criterion or practice to a candidate (e.g. the psychometric assessment process).
2. The candidate has a protected characteristic according to: age; disability; gender identity and gender reassignment; marriage or civil partnership; pregnancy and maternity; race; religion or belief; sex or sexual orientation.
3. The employer also applies the provision, criterion or practice to people who do not share the candidate's protected characteristic.
4. The provision, criterion or practice puts or would put persons with whom the candidate shares the protected characteristic at a particular disadvantage when compared to others (e.g. candidates from a particular ethnic group tend to score lower on an assessment than other candidates).
5. The provision, criterion or practice puts or would put the candidate to that disadvantage (e.g. the candidate fails the assessment process).
6. The employer cannot show the provision, criterion or practice to be a proportionate means of achieving a legitimate aim.

In some cases, an employer that acts in an ostensibly discriminatory manner (e.g. using psychometric assessment methods that have a higher pass rate for some groups than others) will avoid a finding of discrimination by showing that its actions were a proportionate means of achieving a legitimate aim. This is known as 'objective justification'. To be proportionate, a measure has to be both an appropriate means of achieving the legitimate aim and reasonably necessary is order to do so.

If a case is brought, the burden is on the employer to prove justification, and it is for an employment tribunal to undertake a 'fair and detailed analysis of the working practices and business considerations involved' so as to reach its own decision as to whether the treatment was justified.

Significantly lower percentages of Black and Asian candidates pass the current assessment process compared to White candidates. RSSB propose to replace the current assessment process with the new process that is recommended in the T948 report.

Due to the data available, the analysis presented in report T948 only considers age, gender and race as potential protected characteristics.

Questions

1. **Is the process and methodology that RSSB has used to develop a new train driver selection procedure sufficiently objective to ensure that it complies with current legislation and consequently is fair and legally defensible? i.e. with regards to existing case law and current guidance (e.g. Jootley et al. v British Railways Board also known as the 'Paddington Guards Case'; and see also 'A Fair Test' published by the Commission for Racial Equality)?**

   No guarantees can be given in relation to the process given the vagaries of the Employment Tribunal system and the fact that each round of driver assessments will generate fresh data which could indicate discriminatory outcomes. Within that caveat, we can say that the procedure is objective and fair and would stand a very high chance of being successfully defended.

2. **Does the current Driver Selection Governance Group strategy provide a robust method of monitoring and reviewing the process in future from a legal perspective?**

   Yes

3. **Are there any unfair direct or indirect discrimination issues with the recommended tests or overall process considering the evidence of between group differences within the tests and on the job performance?**

   There are no direct discrimination issues with the recommended psychometric assessment process provided it is applied in a standardised way equally to all candidates.

   In terms of indirect discrimination, there do not appear to be any significant issues regarding age or gender.

   The new recommended process is a considerable improvement to the current process in terms of discrimination because it should result in a smaller pass rate difference between whites and non-whites, thus reducing the risk of claims.

   However non-whites are still expected to have a lower pass rate than whites for some assessment methods. According to the trial data presented in Annex 3, Appendix I of the

T948 report, the assessment methods that are likely to have a lower pass rate for non-whites are 2HAND, TRP1, TRP2 and Group Bourdon. These differences could be considered to represent 'particular disadvantage'. Assuming that this is actually the case when the assessment methods are applied as part of recruitment, the key issue is justification for the use of that particular assessment method. This issue is split into two questions:

- **Can the employer establish that it is pursuing a legitimate aim?**

  This will be easily met; the aim of ensuring safe train drivers is clearly legitimate.

- **Can the employer establish that the measures taken to achieve that aim were appropriate and proportionate?**

  - 2HAND – It is a statutory requirement to assess hand coordination, therefore a train company would have a very high chance of defeating any claim on the basis that the assessment was appropriate and reasonably necessary.
  - TRP1 – It is a statutory requirement to assess memory. On the available information this appears to be an appropriate and proportionate test
  - TRP2 – It is statutory requirement to assess reasoning. On the available information this appears to be an appropriate and proportionate test
  -
  - Group Bourdon – It is a statutory requirement to assess attention and given the obvious importance of visual and auditory attention a Train Company would have a strong defence that this criterion in the assessment centre was both appropriate and reasonably necessary.

4. **Do you have any other observations on the selection tests from a legal point of view e.g. the guidance provided on relevant legislation?**

Compliance with the so called "4/5ths rule" does not in itself offer a defence to any potential claims for indirect discrimination. The RSSB is quite right to ameliorate those tests showing pass rates with a lower than 80% correlation but it is more than possible that a test with pass rates lower than 95% pass rate could be held to put a group at a particular disadvantage (subject to the defence of objective justification). As acknowledged in the report, the 4/5ths rule has no statutory force in the UK and is merely a rule of thumb.

If there is disparate performance between a group with a protected characteristic, there is no requirement for a claimant to prove that the disparate performance was because of the protected characteristic, merely that there was such disparate performance.

The question of whether a statistical difference in the performance of different groups is sufficient to amount to a particular disadvantage is a question of fact for tribunals to determine. The higher courts have resisted attempts to persuade them to lay down guidance on this issue, which depends heavily on the context in which the claim is brought.

Winckworth Sherwood

In a claim for indirect discrimination, the trial data from T948 would be of assistance to a claimant. However, the tribunal would also want to access the pass rate data for the specific employer's assessment centre candidates and possibly the pass rate data held by RSSB as part of its on-going evaluation of the new process.

In addition to pass rate data, a tribunal might also take into account disparity in numbers within the candidate pool, which can indicate that the job is not attractive to people in certain groups.

In relation to age, the Equality Act 2010 provides that people will share the protected characteristic of age when they are in the same 'age group'. An age group is defined widely as 'a group of persons defined by reference to age, whether by reference to a particular are or a range of ages'. This definition gives a claimant a good deal of choice when identifying the age group that has allegedly been disadvantaged. The analysis in T948 was only able to compare two age groups (up to 50 years old and 51 or older). In future comparison exercises the RSSB should look at disparate performance by more age groups (e.g. under 25, 26 – 35, 36 – 45, 46 – 55, 55+).

The written communication test (WCT) is designed to measure the driver's ability to communicate effectively in writing. There is a significant discrepancy between white and non-which and between younger and older workers. Written communication is not considered to be safety critical and it is not intended for the WCT to be used a part of the elimination process at the assessment centre. The WCT is appropriate for use as an identifier of training needs but must not be used for other purposes such as tie-breaker or in any way that uses it as part of a decision about whether to eliminate a candidate because it is not necessary to do so.

Winckworth Sherwood LLP
22 January 2013

RSSB Research Programme
Block 2 Angel Square
1 Torrens Street
London
EC1V 1NY

enquirydesk@rssb.co.uk

www.rssb.co.uk/research/Pages/default.aspx