EUROPEAN ORGANISATION FOR THE SAFETY OF AIR NAVIGATION

EUROCONTROL

Review of Workload Measurement, Analysis and Interpretation Methods

CARE-Integra-TRS-130-02-WP2

Edition	**:**	**1.0**
Edition Date	**:**	**19/03/03**
Status	**:**	**Final**
Class	**:**	**CARE**

EUROPEAN AIR TRAFFIC MANAGEMENT PROGRAMME

DOCUMENT IDENTIFICATION SHEET

DOCUMENT DESCRIPTION

Document Title

Review of Workload Measurement, Analysis and Interpretation Methods

EWP DELIVERABLE REFERENCE NUMBER:

PROGRAMME REFERENCE INDEX:	**EDITION:**	1.0
CARE-Integra-TRS-130-02-WP2	**EDITION DATE:**	19/03/03

Abstract

This report was prepared as part of a project being conducted under the EUROCONTROL INTEGRA programme. The aim of the project is to derive principles of workload measurement in man-in-the-loop simulations from experience in non-air traffic management (ATM) domains. This report describes the outcome of Work Package 2. Types of workload measure — performance-based, subjective, and physiological/biochemical — are critically reviewed, and advice is given on methods of selecting the best set of measures for ATM simulations. In the next Work Package, the development of sound experimental designs incorporating these measures will be considered.

Keywords

Workload	Performance	Subjective	Physiology
Air traffic control	Guidelines	Rating scales	Primary task
ATM system	Human factors	Simulation	Secondary task

CONTACT PERSON: Rod Gingell	**TEL:** 00 322 729 3183	**DIVISION:** DIS/ATD

AUTHORS: Eric Farmer & Adam Brownson, QinetiQ

DOCUMENT STATUS AND TYPE

STATUS		CATEGORY		CLASSIFICATION	
Working Draft	☐	Executive Task	☐	General Public	☐
Draft	☐	Specialist Task	☐	EATMP	☑
Proposed Issue	☐	Lower Layer Task	☑		
Released Issue	☑				

ELECTRONIC BACKUP

INTERNAL REFERENCE NAME:

HOST SYSTEM	MEDIA	SOFTWARE
Microsoft Windows	Type: Hard disk	
	Media Identification:	

DOCUMENT APPROVAL

The following table identifies all management authorities who have successively approved the present issue of this document.

AUTHORITY	NAME AND SIGNATURE	DATE
DIS/ATD	Rod Gingell	

DOCUMENT CHANGE RECORD

The following table records the complete history of the successive editions of the present document.

EDITION	DATE	REASON FOR CHANGE	SECTIONS PAGES AFFECTED
A	16/01/2003	First issue of draft	
1.0	19/03/2003	Final version	No further changes

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This report was prepared as part of a project being conducted under the EUROCONTROL INTEGRA programme. The aim of the project is to derive principles of workload measurement in man-in-the-loop simulations from experience in non-air traffic management (ATM) domains. This report describes the outcome of Work Package 2. Types of workload measure — performance-based, subjective, and physiological/biochemical — are described, and examples of each are critically evaluated.

Performance-based: This category can be sub-divided into primary-task and secondary-task measures. The primary task is the task whose workload is being investigated, whereas a secondary task is an additional task used to determine the operator's spare capacity. Examples of these measures are given. Unlike other types of workload measure, there are no standardised performance-based measures. However, guidance is given concerning the development of measures appropriate for various types of task. Objective measures are favoured, but methods of obtaining ratings from subject-matter experts are also discussed.

Subjective: This category includes both uni-dimensional and multi-dimensional scales. The latter, such as the NASA Task Load Index, provide information on the source of excessive workload. Several subjective methods likely to be suitable for the INTEGRA project are identified.

Physiological/biochemical: Some of these measures, such as the pupillary response, are based on the assumption that the individual's level of arousal will vary as a function of workload. In general, these measures are less convenient to use than performance and subjective measures, but they can provide useful additional information (e.g., whether increasing workload produces physiological evidence of stress).

It is recommended that objective measures be used where possible, although well-constructed subjective techniques can yield useful information on the perceived level of effort associated with the operation of particular systems. Qualitative assessment is not recommended for INTEGRA, since more powerful quantitative methods are readily available.

Workload studies may be compromised if operators choose to perform unnecessary tasks (for example, if they manually perform tasks that are automated, owing to lack of trust in the system). Methods of allowing for this possibility are described.

The most appropriate battery of workload measures will depend upon the nature of the tasks and the objectives of the simulation exercise. The report includes a guide to the selection of workload measures for particular applications.

1. INTRODUCTION

1.1 The INTEGRA Project

The INTEGRA Project forms part of the Initiative for Co-operative Actions for R&D in EUROCONTROL (CARE). Its aim is to quantify the benefits of automated support tools for air traffic management (ATM) in real or future environments, against criteria of safety, capacity, economy and environmental impact. The project comprises three phases (Initiation, Definition and Execution). In the Initiation phase, it was found that many real-time human-in-the-loop ATM simulations did not provide quantitative data that would enable INTEGRA to meet its objectives; evidence relating to the criterion of *capacity* was particularly lacking.

This report has been prepared as part of a project ? 'Measuring Workload in Man-in-the-Loop Simulations: Principles Derived from Non-ATM Domains' ? initiated in response to the issues described above (Eurocontrol INTEGRA Task Requirement Sheet No. TRS/130/02). The project is a component of the INTEGRA Definition phase, in which the experimental environment for the simulation/validation exercises planned for the Execution Phase will be specified.

The project is being conducted by QinetiQ and is divided into four Work Packages (WPs):

- WP1: Review of non-ATM human-in-the-loop simulations
- WP2: Review of measurement, analysis and interpretation methods
- WP3: Experimental design and analysis techniques for human-in-the-loop experiments
- WP4: Worked Example

This report describes the outcome of WP2.

1.2 Definitions of workload

"Workload is not an inherent property, but rather it emerges from the interaction between the requirements of a task, the circumstances under which it is performed, and the skills, behaviours, and perceptions of the operator." (Hart & Staveland, 1988).

Workload is sometimes defined operationally in terms of factors such as the task requirements or the effort that must be expended to perform the task. However, it is unwise to consider only one aspect of workload, since these factors inter-relate in complex ways.

Task demands: It might be argued that workload can be inferred from analysis of the tasks required of the human operator. However, individual differences

must be taken into account. For example, a novice and an expert will clearly experience different levels of workload when performing the same task. Skill development produces both economy of action and automatised 'motor programs' that do not require conscious effort. The experienced driver may not even be aware of having changed gear, since this behaviour has been delegated to a motor program.

Effort: The amount of effort expended on a task (i.e., the conscious allocation of mental processing resources) probably corresponds most closely to intuitive notions of the nature of workload. However, when subjected to increased task demands, the individual may choose not to increase the level of effort. Under these circumstances, performance on the task may decline, but recording only measures of effort would lead to the misleading conclusion that workload had not changed.

Performance: Most studies of workload are ultimately concerned with the level of performance that can be achieved. However, performance measures alone cannot serve as an adequate metric of workload. For example, the individual may compensate for increased demand by increasing the level of effort to maintain performance. However, the 'spare capacity' to respond to unforeseen events may be severely reduced, an effect that would not be apparent in the primary-task performance measure.

(see Gartner & Murphy, 1979)

For the reasons described above, it is usually necessary to select a battery of workload measures in any experimental evaluation. Possible measures are described in the next section.

2. INTRODUCTION TO WORKLOAD MEASUREMENT

2.1 Application of workload measures

Workload measurement has been applied to a number of military and industrial problems. Many of the common measures, such as the NASA Task Load Index, were developed for use in aviation, particularly studies of aircrew workload, although they have also been applied elsewhere, such as the nuclear and other safety-critical industries.

2.2 Approach

We do not intend in this report to provide a comprehensive account of the many workload measures that have been reported in the literature. Several detailed reviews have been published, and are cited in the text. Our approach is to consider the most commonly used measures, and indicate their strengths and weaknesses in the context of the INTEGRA programme. Our aim is to identify measures with a sound history that can be used with confidence in INTEGRA studies, rather than promising but unproven methods.

2.3 Types of workload measure

There are three major types of workload measure:

- Performance-based: This category can be sub-divided into *primary-task* and *secondary-task* measures. In this context, the primary task is the task whose workload is under consideration, whereas a secondary task is typically one that is artificially added to determine the amount of 'spare mental capacity'[1] available when the operator is performing the primary task. The rationale is that performance on the secondary task will decline as a function of the demands of the primary task. This perhaps over-simplistic notion is represented schematically in Figure 1. The figure shows variation over time in primary-task demands; the individual's mental capacity is also likely to vary owing to factors such as fatigue.

[1] Alternatively, an 'embedded' task can be used, which can more realistically be combined with the primary task

Advantages

- Primary-task measures provide a direct indication of performance on the task of interest
- Secondary-task performance provides a useful index of spare capacity (e.g., the ability to respond to emergencies or unforeseen events)

Disadvantages

- Performance on the primary task may be insensitive to workload change if operators compensate by increased effort
- Some secondary tasks may be insensitive to the demands of the primary task (e.g., a verbal secondary task may not interfere with a spatial primary task, even if the primary task is very demanding)

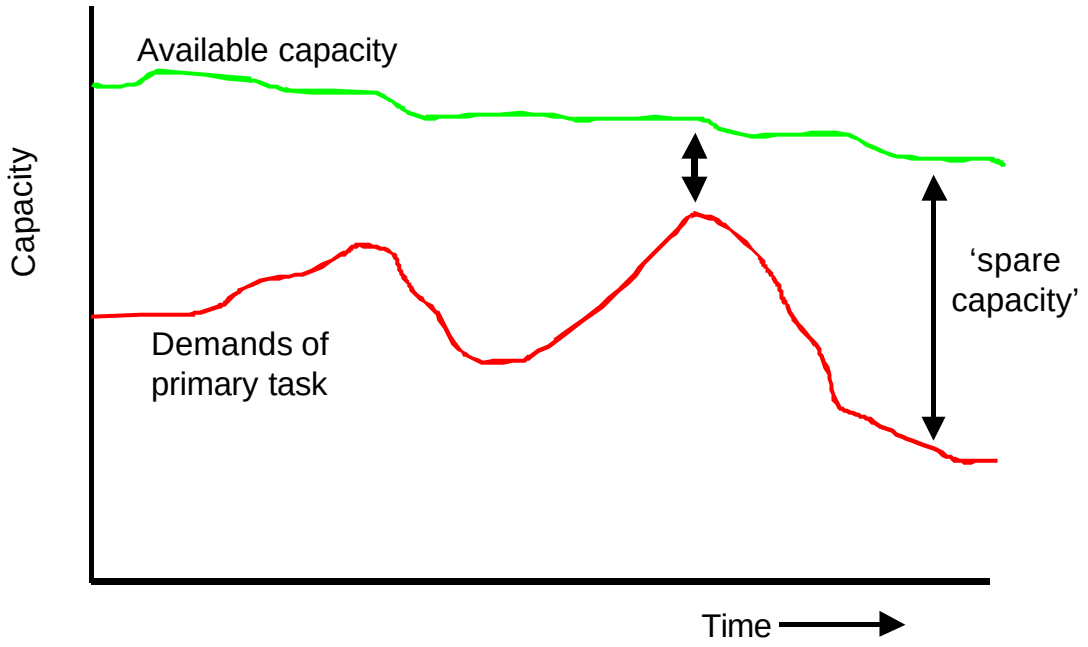


Figure 1. Using secondary tasks to measure spare capacity. It is assumed that performance on the secondary task will reflect the spare capacity available during execution of the primary task

- Subjective: This category includes both uni-dimensional and multi-dimensional scales. The former attempt to elicit a measure of overall workload, whereas the latter address individual components of workload and hence are potentially of some diagnostic value in determining the source of any workload problem.

Advantages
- High 'face validity' and hence acceptance by operators
- Most operators find it fairly easy to assign ratings

Disadvantages
- Operators can rate changing demands of a given task, but find it difficult to compare workload on qualitatively different types of task
- No unanimous agreement on the nature of the components of workload, and hence the set of scales that should be used

- Physiological/biochemical: These measures are based on the premise that workload will induce bodily changes. Some are related to the concept of *arousal*, a continuum that extends from deep sleep to a state of frantic excitement, and appears to be determined by the activity of the reticular formation of the brain. An operator who is overloaded may experience increased arousal, manifested in changes such as increase in heart rate and skin conductance. Since workload is commonly considered to be a stressor, biochemical changes associated with stress ? such as increase in cortisol excretion ? are sometimes assessed in workload studies.

Advantages
- Can often be recorded continuously, with little intrusion on work activities
- Are not affected by biases that sometimes contaminate subjective measures

Disadvantages
- Often a large volume of data is collected, requiring sophisticated analysis
- There may not be a simple relationship between physiological change and performance

3. PERFORMANCE-BASED WORKLOAD MEASURES

3.1 Examples of primary-task performance measures

The appropriateness of particular performance measures is determined largely by the nature of the task. For tasks comprising discrete events to which well-defined responses are required, common measures are:

- reaction time (RT): the time between presentation of a stimulus and execution of a response. For tasks in which RTs are no longer than a few seconds, it is customary to record RT to the nearest millisecond
- accuracy: often expressed in the form of percentage or proportion of errors

It is usually essential to record both types of measure, since a speed/accuracy trade-off is a feature of many tasks. The operator can choose, for example, to increase speed at the expense of accuracy (e.g., Fitts, 1966); if only reaction time measures were recorded, it might erroneously be concluded that performance had improved.

A difficulty of interpretation sometimes emerges when speed and error measures trade off. Without knowledge of the precise form of the trade-off function, it is difficult to determine whether performance has simply moved to a different point on the same speed/accuracy function or has genuinely improved or declined. We recommend a pragmatic approach: in safety-critical industries such as ATM, a substantial decrease in accuracy should give concern even if speed is found to improve.

In practice, speed and accuracy measures are often consistent. For example, a poor system interface will tend to increase reaction time and decrease accuracy, and practice will tend to improve both speed and accuracy. Figure 2 shows results reported by Farmer et al (2000) for training on a threat assessment task; a clear pattern of improvement is apparent for each measure.

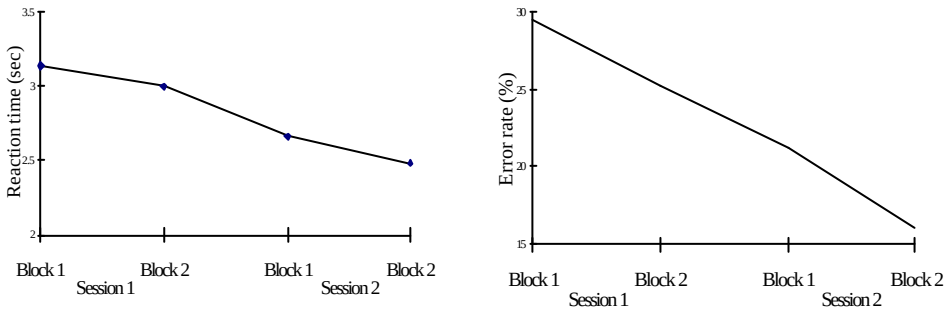


Figure 2: Improvement in performance with practice on a threat assessment task

For tasks comprising continuous movement (e.g., steering), a commonly used measure is:

- root mean square (RMS) error: Error, in the form of the distance between actual and desired position, is recorded at a suitable rate (e.g., 10 Hz). The use of RMS error rather than the arithmetic mean penalises inconsistency. For example, the following sets of error values:
  2 ? 4 ? 3
  3 ? 3 ? 3
  have the same mean values, but the RMS value for the former is higher.

For vigilance tasks, characterised by the presentation of weak, irregular signals (e.g., radar monitoring), Signal Detection Theory (SDT; e.g., Egan, 1975) is often applied. Responses are coded as shown in Table 1; for example, a 'hit' is recorded when subjects make a positive response to a genuine signal.

		Stimulus	
		No	Yes
Response	No	Correct rejection	Miss
	Yes	False alarm	Hit

Table 1: Stimulus–response outcomes for a detection task

SDT assumes that there is noise on sensory channels (as an analogy, even in a completely dark room, an individual will perceive spots of light). In a signal detection task, the subject must distinguish such noise from genuine, very weak, signals. Since noise is random, it can be represented as a Gaussian distribution. The application of SDT yields two major measures:

- d? ('d prime'), which is a measure of the discriminability of the signal from noise, and corresponds to the distance between the means of the noise and signal plus noise distributions
- ß, which is a measure of the individual's response criterion, i.e., the amount of sensory evidence required before a decision is made that a signal has been presented (with a strict criterion, the individual will commit few false alarms but is likely to miss genuine signals; with a lax criterion, the individual will incur few misses but commit a considerable number of false alarms).

In practice, it is not necessary to compute these scores directly from experimental data. Having recorded the proportion of hits and false alarms, the experimenter can consult published tables of d? and ß values. SDT is of considerable value in safety-critical applications, since it indicates whether

poor performance is attributable to difficulty in detecting events or to an inappropriate response criterion. For example, where all signals must be detected and there is little penalty for false alarms, a very strict criterion should not be adopted.

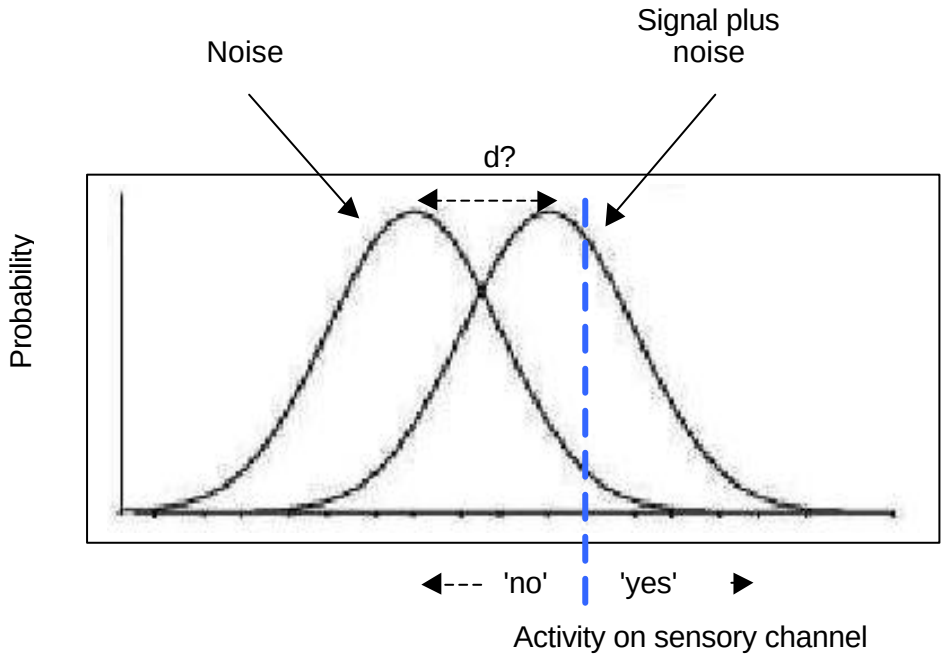


Figure 3. Signal Detection Theory: Effect of signal presentation on sensory activity. The blue line depicts the response criterion selected by the subject (the amount of evidence required for a 'yes' decision)

Since most studies of the type envisaged under the INTEGRA project will be concerned with overall system performance, we strongly recommend that measures of performance are always included, even if they do not provide a definitive workload profile in isolation. Where objective measures such as those described above are not available, ratings of performance by subject-matter experts (SMEs) should be obtained.

Subjective ratings of performance have a number of possible shortcomings. However, most of these can be overcome by careful design of the scoring system. For example, scorers are often reluctant to use the extremes of rating scales, with the result that sensitivity to individual differences is compromised. However, the use of behaviourally anchored rating scales (BARS), in which each point on the scale is assigned a verbal description, greatly improves the allocation of ratings. Figure 4 from a QinetiQ study (Weston-Lovelock & Abram, 1996), illustrates the benefits of the use of BARS.

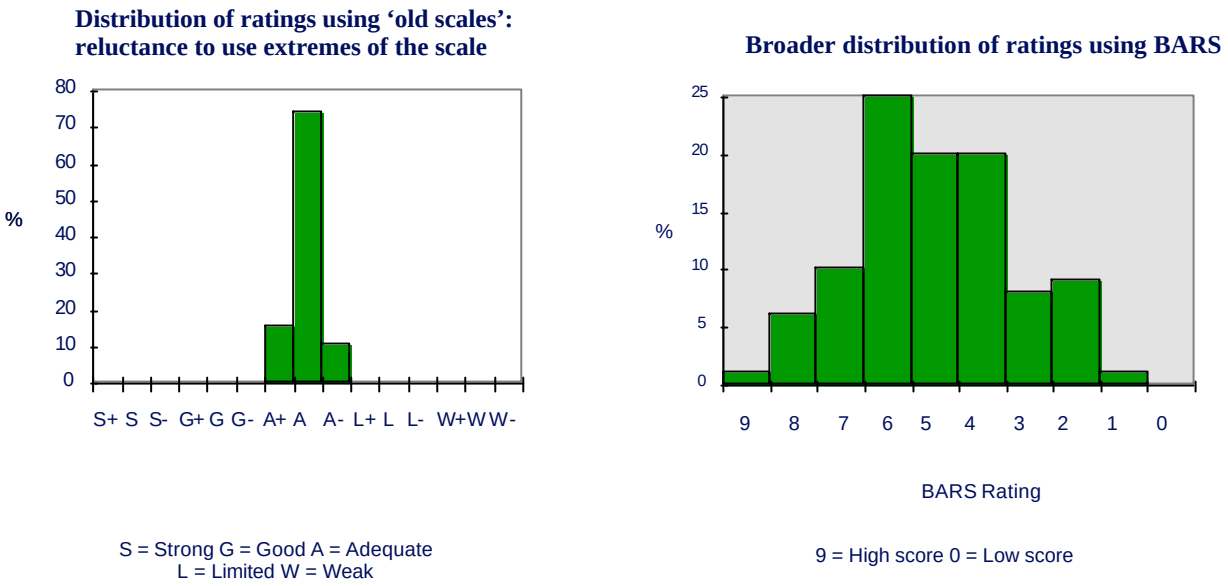


Figure 4. Effect of using BARS on distribution of scores (from Weston-Lovelock & Abram, 1996)

3.2 Secondary task measures

Secondary task measures that have been used include:

Interval production: The individual is asked to tap at a specified rate. As workload increases, the intervals between taps become increasingly variable.

Time estimation: The individual is asked to estimate how much time has elapsed (for example, since the start of a simulation session). In general, time intervals are progressively under-estimated as workload increases (e.g. Casali & Wierwille, 1983).

Random number generation: as workload increases, the individual may resort to well-learned sequences (e.g., "1, 2, 3"). This decrease in randomness can be quantified mathematically (e.g. Zeitlin & Finkelman, 1975).

Probe reaction time: A stimulus unrelated to the primary task appears periodically; RT to this stimulus is assumed to reflect the demands of the primary task.

The above tasks can yield information on the overall demands of the primary task. However, under some circumstances secondary tasks can provide diagnostic information:

Memory search: In the memory search task developed by Sternberg (e.g., 1969), the subject is presented with a 'memory set' of items, which is then removed and replaced after an interval by a 'probe' item. The task is to indicate whether the probe item was present in or absent from the memory set. The typical result for 'yes' responses is shown in Figure 5a. There is a linear relationship between memory set size and reaction time, the slope of the function suggesting that individuals access material in short-term memory serially, at a rate of about 25/sec. The y-intercept of the function is assumed to reflect the 'fixed costs' associated with acquiring the stimulus material and executing a response. Figures 5b and 5c reflect hypothetical findings in which the slope and intercept, respectively, are affected by increased workload on the primary task. (In statistical terms, there is an interaction between memory set size and workload in Figure 5b, but only main effects of these factors in Figure 5c.) The former result suggests an increase in the cognitive demands of the primary task (since the rate of memory search has slowed); the latter suggests that input/output demands have been increased.

Measuring distraction: the Peripheral Detection Task

Martens M.H. & van Winsum, W. (2000). Internal Report. TNO Human Factors, Soesterberg, The Netherlands

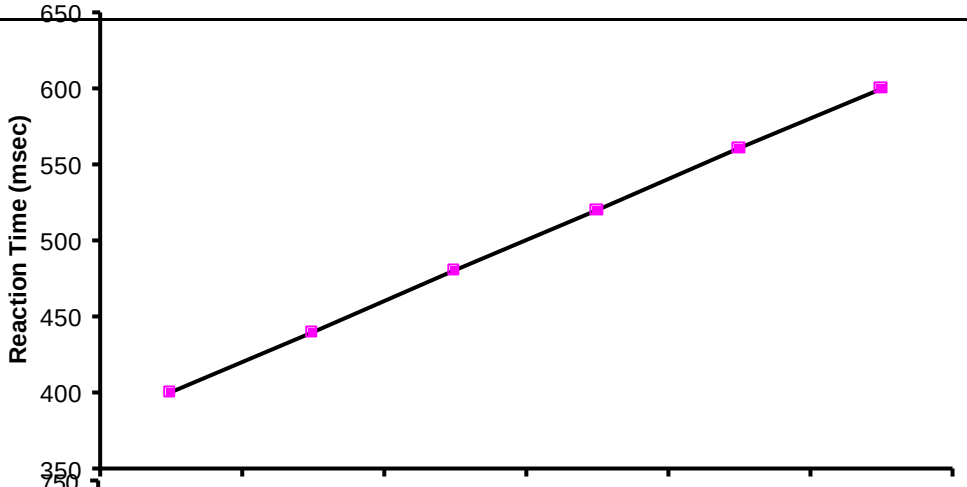
The impact of in-vehicle systems on driver distraction and workload were evaluated using a Peripheral Detection Task (PDT), similar to the probe reaction time methodology. Subjects performed a driving task in a simulator, and were required to respond as soon as a red square was detected on the simulator screen. Stimuli were presented at varying locations within the driver's visual field to investigate the effect of 'Cognitive Tunnelling' during periods of increased task difficulty. Reaction time data and signal detection data demonstrated the high sensitivity of the PDT to peaks in workload resulting from both increased driving task difficulty and the demands of the in-vehicle system.

3.2.1 Selecting a secondary-task measure

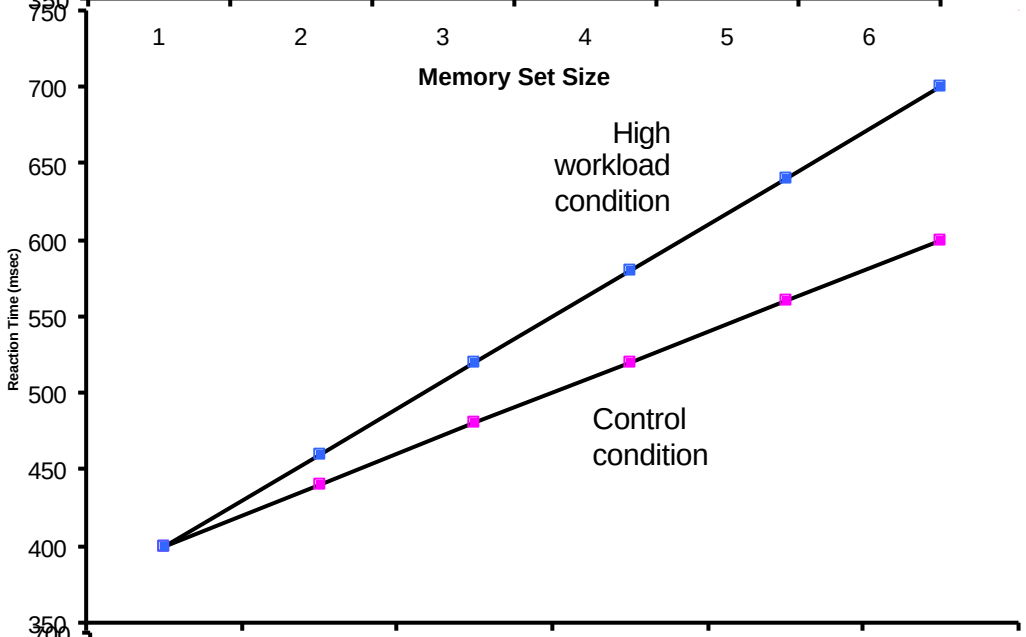
Secondary task measures must be selected with care. The rationale for using a secondary task is that it will compete for the finite ‘mental resources’ required to perform the primary task. However, there is evidence for the existence of discrete resource pools, analogous to the specialised processors

found in computer systems. For example, Farmer et al (1986) showed that a simple verbal task (repeatedly saying aloud the sequence “1, 2, 3, 4”) interfered with verbal but not spatial reasoning, whereas a simple spatial task (repeatedly tapping four large metal plates in sequence) produced the opposite pattern of interference. These findings supported the notion of distinct verbal and spatial sub-systems within human working memory. The idea of ‘multiple resources’ has been championed by Wickens (e.g., 1992). Although the Wickens model has been subject to criticism (e.g., Jones & Farmer, 2001), it is clear that some combinations of tasks will mutually interfere to a greater extent than others. A secondary task should therefore be chosen that is likely to interfere with the primary task. For example, if the primary task is primarily spatial in nature, the secondary task should also require spatial processing; performance of a verbal secondary task may be insensitive to changing demands of the primary task, since it is not in contention with this task.

a



b



c

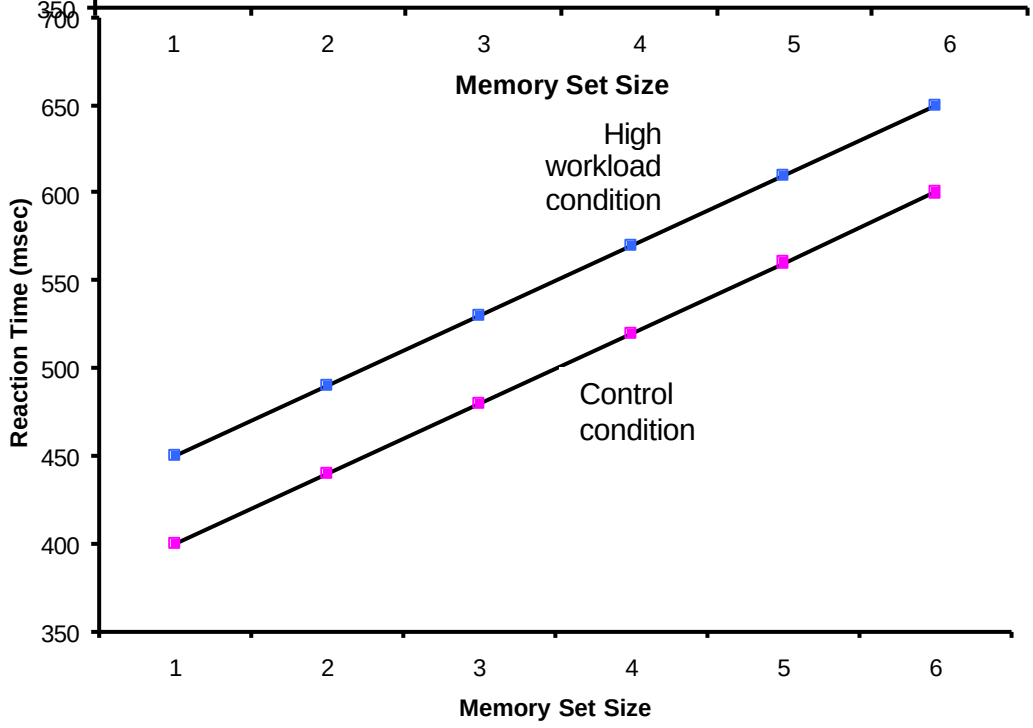


Figure 5. Hypothetical memory search functions

3.2.2 Example of secondary task methodology

Primary Task Disruption from Multiple In-Vehicle Systems

Lansdown T.C., Brook-Carter N. & Kersloot T. (2002). *ITS Journal*, 7(2): 151-168

Secondary task methodology was employed to explore the effects of presenting visual and auditory stimuli, simulating in-vehicle systems such as navigation, concurrently with a simulated driving task. The secondary tasks were found to interfere with the execution of the primary task, although in this experiment no differences were found between the three experimental conditions (visual, auditory, and both visual and auditory). Subjective data showed that overall mental workload and anger and frustration were reported as being higher during the dual-task conditions. The authors concluded that:

- secondary tasks interfere with performance on the primary task;
- primary task measures do not indicate the operator's level of effort;
- subjective measures are sensitive to the operator's own perceptions of the workload, providing data that could not otherwise be obtained.

4. SUBJECTIVE WORKLOAD MEASURES

A variety of subjective measures have been developed, and applied in many studies, particularly those of the flight deck. The best-established measures are described below.

4.1 Modified Cooper-Harper scale

This method is based upon a scale that was originally developed to allow test pilots to rate the handling qualities of aircraft. It uses a decision tree to arrive at a value of 1–10.

Reliability	Dependent on operator acceptance; generally high reliability
Validity	Highly correlated with task performance and other workload measures
Sensitivity	Sensitive to variations in task difficulty. Monotonic relationship with loading level
Diagnosticity	Not diagnostic
Practicality	Easily administered. Decision tree-based, but tree only initially consulted. Subject provides a rating between 1 and 10
Intrusiveness	Low intrusiveness
Summary	Useful as a simple measure of overall workload, but not diagnostic

4.2 ISA

Instantaneous Self Assessment (ISA) was developed as a simple tool with which an operator can estimate perceived workload during real-time simulated or actual tasks. The operator is required to give a rating of perceived workload or 'busyness' on a scale from 1 (very low) to 5 (very high).

Reliability	Reliable if operators report ratings accurately
Validity	Highly correlated with NASA TLX and other workload measures
Sensitivity	Moderate sensitivity, limited by 5-point rating scale
Diagnosticity	Not diagnostic
Practicality	Very practical measure, with minimal equipment required
Intrusiveness	Lowest intrusiveness of all subjective measures
Summary	Useful for a rough indication of workload, particularly in dynamic environments such as the flight deck

4.3 Subjective Workload Assessment Technique (SWAT)

This instrument was developed by Gary Reid at Wright-Patterson Air Force Base, and has been used in a number of aircrew studies. It includes scales for time load, mental effort load, and psychological stress load, each scale having three levels. Subjects are asked to rank the 27 possible combinations of levels on the three scales before providing ratings for particular tasks or events.

Reliability	Reliable, even if reporting delayed by up to 30 min
Validity	Underlying dimensions not empirically validated
Sensitivity	Generally less sensitive than TLX, but still high and extensively tested
Diagnosticity	Multidimensional: three scales (time load, mental effort load, psychological stress load). These have been found to be differentially diagnostic (Moroney, p.101)
Practicality	Two steps: scale development (scale ranking of 27 scale combinations); event scoring.
Intrusiveness	Slightly more demanding than other subjective measures (Step 1 can take up to 45 min (Wierwille & Eggemeier, p.268). Lower user acceptance than TLX. Completed off-line
Summary	A test that provides useful data, but requires more effort than some other multi-dimensional scales and therefore not recommended in the present context

4.4 NASA Task Load Index (TLX)

The NASA TLX was developed by Hart and her colleagues, and has been used in many aviation studies.

Reliability	High reliability
Validity	Extensively validated
Sensitivity	High sensitivity
Diagnosticity	Multidimensional: six scales (Mental, Physical, Temporal, Effort, Performance, Frustration); differentially diagnostic; global score can also be calculated
Practicality	Two steps: event scoring and paired comparison weighting process to determine the importance of each factor for the task in question (the latter not necessarily required)
Intrusiveness	Takes 1 or 2 minutes to complete, off-line
Summary	A very well-established test with a sound basis, which could be used in INTEGRA studies

Age differences in perceived workload across a short vigil **Bunce, D. & Sisa, L. (2002). *Ergonomics*. 45(13): 949-960.**
A demanding, high event rate, vigilance task was used to investigate age differences in perceived workload. Two groups (aged 16–35 and 45–65) took part in the study, which involved a practice session and a test. Primary task data showed a significant decrement in performance during the course of the 9-minute vigilance task, but this effect was not linked to age. However, subjective workload data (NASA TLX) showed that older participants reported a significantly greater increase in workload from the practice session to the test session than was reported by the younger group of participants.

4.5 DRA Workload Scales (DRAWS)

These scales were developed by the Defence Research Agency (DRA), a predecessor of QinetiQ. They were based on factor analysis of a large body of performance and workload data, and were subjected to validation studies during their development.

Reliability	Good reliability established
Validity	Validated
Sensitivity	Shown to be sensitive in several studies
Diagnosticity	Provides information on the four dimensions of workload recovered from empirical data: input demand, central demand, output demand, and time pressure
Practicality	Easy to administer: only four ratings required
Intrusiveness	Takes about two minutes; less after familiarisation
Summary	Based on a slightly different factor structure from the TLX, but has similar properties. Like TLX, suitable for INTEGRA studies

5. PHYSIOLOGICAL/BIOCHEMICAL MEASURES

5.1 Event-Related Potentials (ERPs): P300

ERPs are measures of the brain activity that follows presentation of a signal. The P300 (so named because it is a positive electrical change that occurs about 300 milliseconds after stimulus presentation) does appear to respond to workload (Isreal, Wickens, Chesney & Donchin, 1980). However, this method is probably best reserved for very detailed basic research on workload.

Reliability	The effect appears to be reproducible
Validity	Amplitude of the P300 has been shown to reduce with increasing workload
Sensitivity	Probably very sensitive, since the method is based upon precise measurement of aspects of brain activity
Diagnosticity	P300 amplitude is sensitive to increase in task difficulty. Also sensitive to motor artefacts.
Practicality	Difficult to analyse data; high noise-to-signal ratio; requires calibrating to each individual; considerable and costly supporting instrumentation and computer equipment required; requires trained personnel to supervise.
Intrusiveness	EEG electrodes positioned on scalp
Summary	Not recommended for INTEGRA

5.2 Heart rate

Heart rate has been used in a variety of studies, including those of flight deck operations. It is easy to record, but its major weakness is that it is affected by factors other than workload, such as anxiety and physical activity.

Reliability	Effects appear to be replicable
Validity	Systematic relationships between HR and a variety of information-processing activities in laboratory and field environments (Harris et al, 1989; Wierwille & Connor, 1983). Although HR varies with workload, it is affected by a number of other factors; hence, the validity of this measure in particular studies may be questionable
Sensitivity	Sensitive to variations in task demand, but also sensitive to contamination from physical effort, emotions and stress; large signal/noise ratio (Kramer, 1991)
Diagnosticity	Very limited diagnosticity. Overall index of general arousal or physical work. Integrated index of overall effect of task demands and the emotional response of the operator to them (Hart & Wickens, 1990)
Practicality	Portable recording equipment is available. Data susceptible to contamination from physical effort, emotions and stress. Data require sophisticated analysis and the relationship with performance may be complex
Intrusiveness	Low intrusiveness: electrodes do not need precise location.
Summary	Vulnerable to confounding effects of other variables; however, if recording heart rate variability (see below), it is worthwhile also to analyse heart rate

An analysis of mental workload in pilots during flight using multiple psycho-physiological measures

Wilson, G.F. (2002). *The International Journal of Aviation Psychology*, 1: 3-18.

Ten pilots were required to fly a 90-minute scenario in an experiment to test the reliability of psycho-physiological measures of workload. Each pilot performed the same scenario twice to assess the test-retest reliability of the measures. Cardiac, electrodermal and electrical brain activity measures were highly correlated and exhibited changes in response to the demands of the flights. Heart rate was more sensitive than heart rate variability.

5.3 Heart rate variability (HRV)

Early attempts to relate heart rate variability to task demands were largely unsuccessful. However, it was later found that the power at 0.1Hz, determined by spectral decomposition of the HRV data, was a good measure of mental effort (Aasman, Mulder & Mulder, 1987).

Reliability	A reliable measure, successfully applied in laboratory and non-laboratory environments
Validity	Generally understood to be a valid measure. However, up to 30 analysis techniques have been identified, not all of which may produce valid results
Sensitivity	High sensitivity, but vulnerable to contamination from stress and ambient environment
Diagnosticity	Global measure of effort
Practicality	One of the most practical physiological measures. Calibration and rest measurements required between test conditions (Papillo & Shapiro, 1990). Since at least two minutes' worth of data must be collected to obtain an estimate of HRV, this measure is not suitable for tasks of very short duration
Intrusiveness	Modern equipment to record cardiac measures is relatively unobtrusive
Summary	If there is interest in obtaining an objective measure of effort for periods of at least a few minutes, HRV can be recommended

Free Flight and Air Traffic Controller Mental Workload

Hilburn, B.G. (1997). Presented at the *Ninth International Symposium on Aviation Psychology*, 28 April–1 May 1997. Columbus, Ohio, USA

Eight air traffic controllers took part in an ATM experiment. Two levels of system automation were tested, with high and low traffic volumes. Measures of efficiency, subjective workload, pupil diameter, and heart rate variability at 0.1 Hz were recorded. Respiration was also recorded, to allow removal of the influence of respiration rate on HRV.

Both pupil diameter and heart rate variability were sensitive to differences in traffic volume and level of automation. HRV was significantly lower where traffic volume was high and the level of automation was low. It was therefore concluded that HRV was applicable and sensitive as a workload measure in ATM environments.

5.4 Pupillary response

Pupillary response has been used in workload studies, the underlying rationale being that arousal will increase as a function of workload. Although this measure has been used in practical settings, the required equipment is fairly intrusive and the recording and analysis process is likely to be costly.

Reliability	Small but reliable differences found; can assess relative workload in different situations
Validity	High validity as global measure of workload
Sensitivity	High sensitivity, but also very sensitive to emotional & environmental factors (lighting etc), and hence requires tight experimental control that may be difficult to maintain in realistic simulation exercises
Diagnosticity	Limited diagnosticity; global screening device
Practicality	Practical only in laboratory setting. Requires costly, but commercially available, equipment.
Intrusiveness	May be intrusive in simulation trials
Summary	A good research tool, but costs outweigh benefits in the context of INTEGRA

Human machine interfaces for ATM: objective and subjective measurements on human interactions with future flight deck and air traffic control systems

Jorna, P. (1997). Internal report. NLR, Amsterdam

In a simulation of ATM, controllers experienced three types of datalink user interfaces under both high and low traffic densities. Amongst other physiological and subjective measures, pupil size was measured, and accurate measurements of small differences were obtained. Pupil size decreased as a function of the level of interface integration, and increased significantly when high traffic density was experienced. Jorna notes that pupil size might have been expected to *decrease* as a function of the number of radar plots on the screen, due to the increased light present. Pupil size is therefore a sensitive measure of mental workload.

5.5 Eye Blink

This measure appears to be sensitive to visual workload.

Reliability	Consistent relationships between task demands and blink latency/duration have been found across a variety of populations and tasks. However, patterns of data for blink rate are less consistent (Kramer, 1991).
Validity	Valid measure of visual workload
Sensitivity	Sensitive to variations in task demand in flight environments (Hughes et al, 1990; Wilson & Fullenkamp, 1991)
Diagnosticity	Diagnostic of visual workload
Practicality	Requires EOG portable equipment; calibration; electrodes. Has been successfully applied in aircraft and simulators (e.g., Wilson et al, 1987).
Intrusiveness	EOG electrodes placed above and below the eyes. Otherwise low intrusiveness, high operator acceptance
Summary	Requires specialised equipment, but of some value if there is particular interest in visual processing demands of a simulation

Physiological workload reactions to increasing levels of task difficulty

Veltman, J.A., and Gaillard, A.W.K. (1998). *Ergonomics*, 41 (5): 656-669

The sensitivity of physiological measures to mental workload was investigated in a flight simulator. Pilots were required to fly through a tunnel with different levels of difficulty, and perform a memory task with four levels of difficulty. Rest periods before and after the experiment were used as a baseline. During the more difficult conditions, the time between successive eye blinks (blink interval) increased, and the blink duration increased as more visual information had to be processed. Increasing the difficulty of the memory task led to a decrement in blink interval; this was probably caused by sub-vocal rehearsal of target letters during memory search. Blink duration was not affected by memory load.

5.6 Cortisol

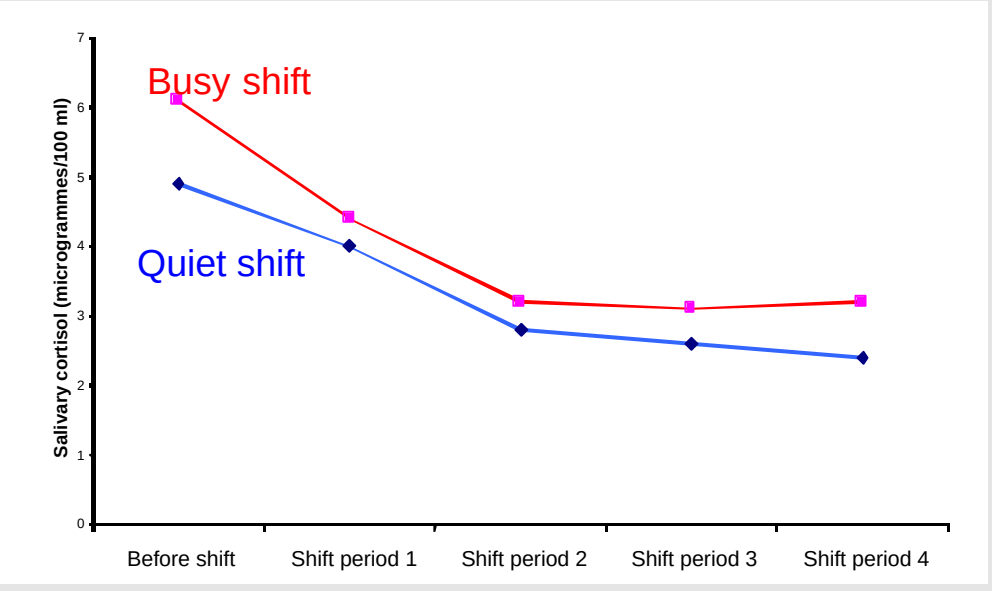
Salivary cortisol excretion provides a measure of the stress that is sometimes associated with high workload. For example, Farmer et al (1991) showed that cortisol excretion among UK air traffic controllers increased during busy periods.

Reliability	Few studies directly addressing reliability
Validity	Has been shown to be associated with a state of stress
Sensitivity	Relatively high sensitivity
Diagnosticity	Provides only an overall measure of stress
Practicality	Requires biochemical analysis, which is unlikely to be available in-house
Intrusiveness	Fairly unobtrusive — saliva sample collected at intervals (e.g., breaks in a simulation session)
Summary	A useful measure if it is suspected that workload will be a significant source of stress. However, this information can typically be obtained more easily by other means

Stress in Air Traffic Control II: Effects of Increased Workload

Farmer, E.W., Belyavin, A.J., Tattersall, A.J., Berry, A. & Hockey, G.R.J. Farnborough: RAF Institute of Aviation Medicine Report No. 701, 1991.

As part of a large-scale study of stress and workload in UK air traffic control, Farmer et al recorded salivary cortisol before and during quiet and busy shifts. There was a clear elevation of cortisol excretion in the high workload condition, represented in the diagram below.



5.7 DC Shift

It has been found that workload has an effect on the DC level of brain electrical activity. Figure 6 shows data reported by Pleydell-Pearce, in a study conducted for QinetiQ.

Reliability	The effect is reproducible
Validity	Good correlation between objective measures of workload and DC level
Sensitivity	Can discriminate between several levels of workload
Diagnosticity	Limited diagnosticity
Practicality	Requires skilled placement of electrodes and data analysis
Intrusiveness	Electrodes may be intrusive in simulation exercises
Summary	An effective, objective measure of workload, but not suitable for INTEGRA

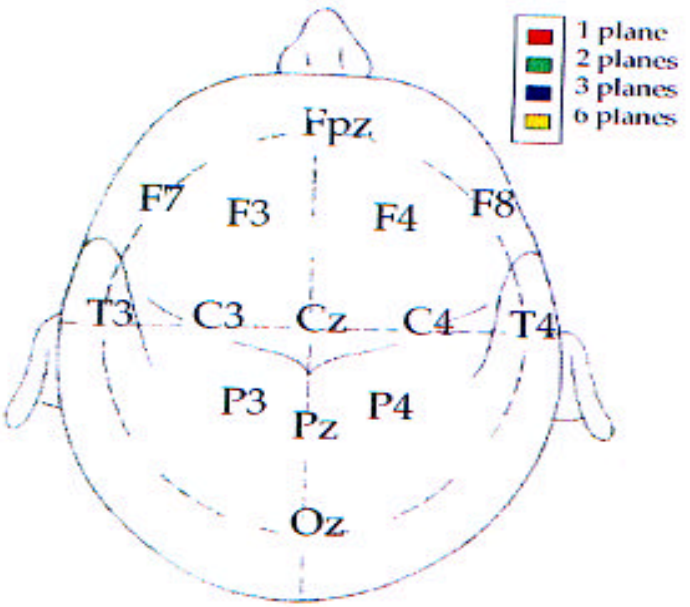


Figure 6a. Placement of electrodes for DC Shift studies (source: Pleydell-Pearce)

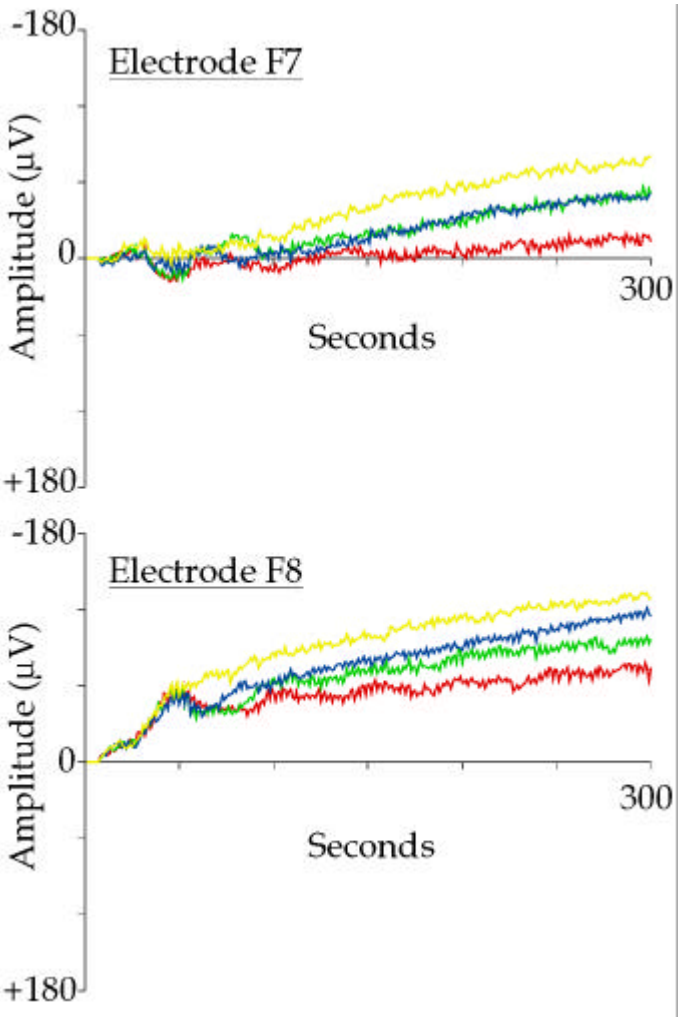


Figure 6b. DC Shift for different levels of workload in an air traffic control task (source: Pleydell-Pearce)

6. FURTHER ISSUES IN WORKLOAD MEASUREMENT

6.1 Subjective versus objective assessment

Although there are useful subjective measures of workload, including some developed in our own laboratories, we recommend that the battery of workload measures selected for any simulation exercise contain at least one objective measure. Where possible, performance of the simulation-based task should be recorded objectively, using measures such as those described earlier. However, subjective workload measures are useful in characterising the perceived demands of the task, which are an important aspect of any system and influence the operator's acceptance of the system.

Dissociations between measures may occur under some circumstances (see, for example, Yeh & Wickens, 1984; Farmer, 1993). For example, increase in workload may lead to a corresponding increase in the effort expended by the operator. Under these conditions, performance may be maintained, but subjective workload considerably increased. Alternatively, the operator may reduce the level of effort expended if the task is perceived to impose such high demands that satisfactory performance is impossible.

For these reasons, a workload profile rather than a single measure should be generated. The profile should include, as a minimum, both the level of performance attained and the effort required.

6.2 Quantitative versus qualitative data

Given the ready availability of quantitative workload measures, there is little reason to rely on qualitative assessment such as 'acceptable/unacceptable'. Even when recourse to subjective measures is necessary, there is overwhelming evidence that such measures can be expressed quantitatively. Qualitative measures are unlikely to reveal the subtle effects on workload associated with complex systems such as ATM.

6.3 Task criticality

There is a possibility that operators faced with a new system such as a decision aid will choose to perform tasks that are in fact unnecessary. This activity will in turn artificially inflate the workload associated with system operation. For example, an operator who has not developed trust in an automated system may perform, or check the outcome of, many tasks that are performed entirely adequately by the system. Under these circumstances, the automated system may actually cause an apparent increase in workload.

Workload measures are unable to detect when an unnecessary task has been performed; they can indicate only the demands of the task as it is performed. We recommend the following steps:

- Perform a 'normative' task analysis for the simulation, based on the intended operation of the system. A good introduction to task analysis can be found in Kirwan and Ainsworth (1992). A further benefit of performing a task analysis is that it may help to indicate segments of the simulation on which workload measurement should focus

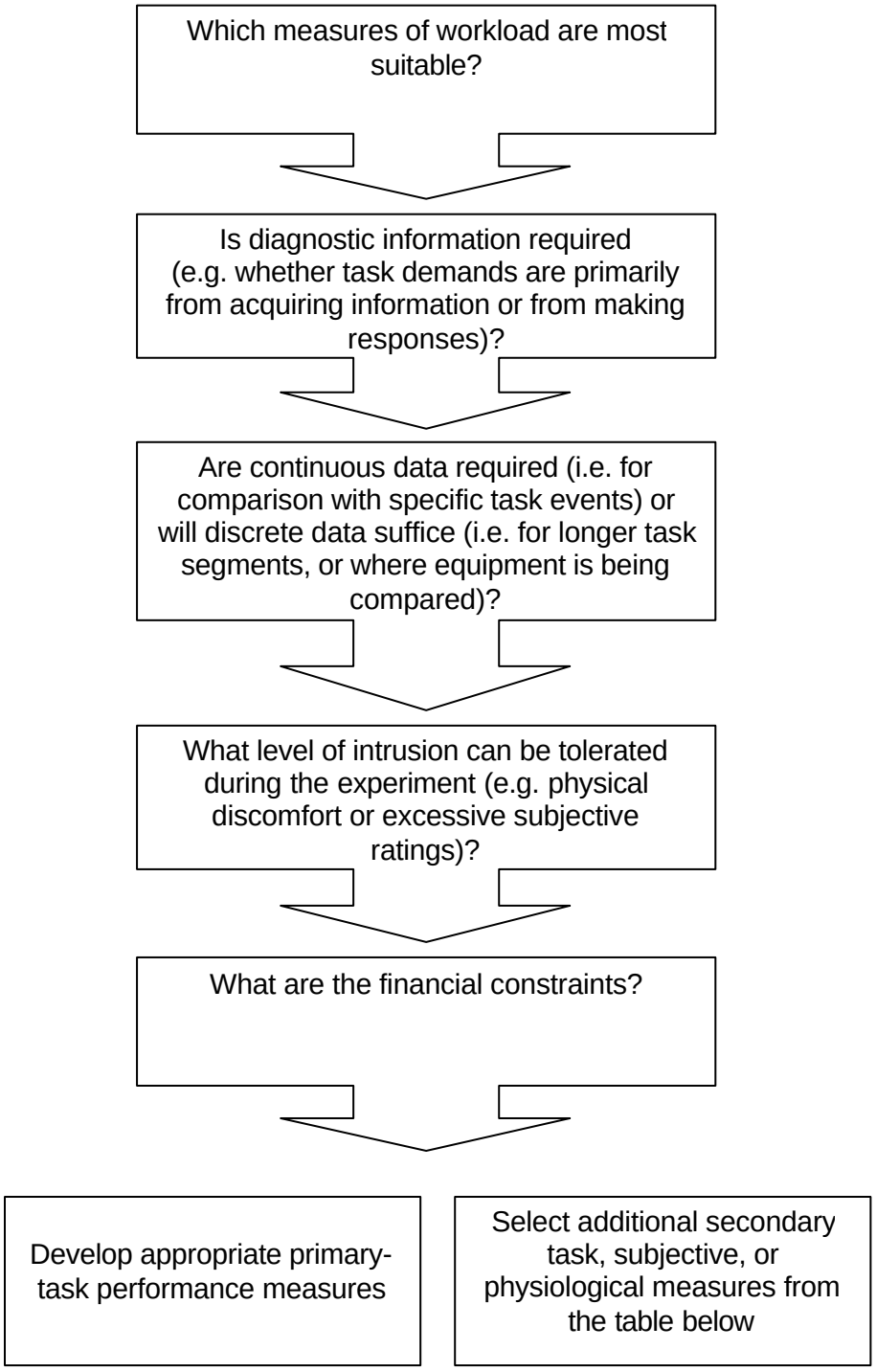
- Provide training for experimental participants, emphasising adherence to the intended procedures and modes of operation embodied in the task analysis

- Instruct the participants to assume that automated sub-systems will perform reliably; participants are more likely to be willing to suspend their mistrust of a system during a simulation than in real operations

- Directly observe (or video-tape for future analysis) the participant's activity during the simulation for evidence that the tasks are being performed as planned

- Relate workload measures to particular events or activities during the simulation; if appropriate, allowance can later be made for unnecessary actions performed by the participant

6.4 Selecting a Set of Workload Measures



Measures recommended for INTEGRA studies

	Diagnosticity	Continuous vs. Discrete	Intrusiveness	Cost
MCH	No	Discrete	Low	Low
ISA	No	Discrete	Low	Low
Secondary Tasks	No	Discrete	Medium	Low
Primary Task	No	Continuous	Low	Low
HR	No	Continuous	Medium	Medium
HRV	No	Continuous	Medium	Medium
NASA-TLX	Yes	Discrete	Low	Low
DRAWS	Yes	Discrete	Low	Low
Blink Rate	Yes	Continuous	Medium	High

Valid measures of workload that are nevertheless not recommended for INTEGRA studies

	Diagnosticity	Continuous vs. Discrete	Intrusiveness	Cost
Cortisol Excretion	No	Discrete	Medium	Low
P300	No	Continuous	High	High
SWAT	Yes	Discrete	Low	Low
Pupillary Response	Yes	Continuous	High	High
DC Shift	No	Continuous	High	High

7. REFERENCES

Aasman, J., Mulder, G., & Mulder L.J.M (1987). Operator effort and the measurement of heart-rate variability. *Human Factors* , 29: 161-170

Casali, J.G. & Wierwille, W.W. (1983). Communications-imposed pilot workload: A comparison of sixteen estimation techniques. *Proceedings of Second Ohio State University Symposium on Aviation Psychology*, 223-235.

Egan, JP (1975). *Signal Detection Theory and ROC Analysis*. New York: Academic Press.

Farmer, E.W. (1993). Conceptual issues in workload. *Proceedings of Conference on Workload Assessment and Aviation Safety*. London: Royal Aeronautical Society.

Farmer, E.W., Berman, J.V.F. & Fletcher, Y.L. (1986). Evidence for a visuo-spatial scratch-pad in working memory. *Quarterly Journal of Experimental Psychology*, 38A, 675-688, 1986. Reprinted in P.T. Smith and R.A. Boakes (Eds), *Human and Animal Memory*. London: Lawrence Erlbaum Associates, 1986.

Farmer, E.W., Jordan, C.S., Belyavin, A.J., Bunting, A.J., Tattersall, A.J. & Jones, D.M. (1995). *Dimensions of operator workload: Final report*. Unclassified DRA Report: DRA/AS/MMI/CR95098/1.

Farmer, E.W., van Rooij, J., Riemersma, J., Jorna, P., & Moraal, J. (2000). *Handbook of Simulator-Based Training*. Aldershot: Ashgate.

Fitts, PM (1966). Cognitive aspects of information processing: III. Set for speed versus accuracy. *Journal of Experimental Psychology*, 71, 849-857.

Gartner, W.B. & Murphy, M.R. (1979). Concepts of workload. In B.O. Hartman & R.E. McKenzie (Eds.), *Survey of Methods to Assess Workload*. AGARDograph No. 246. Neuilly-sur-Seine: AGARD.

Harris, R., Bonadies, G. & Comstock, J.R. (1989). Usefulness of heart measures in flight simulation. *Proceedings of the Third Annual Workshop on Space Operations, Automation and Robotics*. Houston, Texas, NASA Johnson Space Center, cited in A.F. Kramer (1991), Physiological metrics of mental workload: A review of recent progress. In D. Damos. Multiple Task Performance. London: Taylor & Francis.

Hart, S.G. & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock & N. Meshkati (Eds.) *Human Mental Workload*. Elsevier.

Hart, S.G. & Wickens, C.D. (1990). Workload assessment and prediction. In H.R. Booher (Ed.) *Manprint: An approach to systems integration* (pp. 257-296). New York: Van Nostrand Reinhold, cited in W.W. Wierwille & F.T. Eggemeier, Recommendations for mental workload measurement in a test and evaluation environment. Human Factors. 35(2), 263-281.

Hilburn, B.G. (1997). Free Flight and Air Traffic Controller Mental Workload. National Aerospace Laboratory (NLR), Amsterdam. Presented at the *Ninth International Symposium on Aviation Psychology*, 28 April - 1 May. Columbus, Ohio, USA

Hughes, E.R., Hassoun, J.A., Ward, G.F. & Rueb, J.D. (1990). *An assessment of selected workload and situation awareness metrics in a part-mission simulation*. (Technical Report ASD-TR-90-5009). Wright-Patterson Air Force Base, OH: DCS for Integrated Engineering and Technical Management, Aeronautical Systems Division, Air Force Systems Command, cited in W.W. Wierwille & F.T. Eggemeier, Recommendations for mental workload measurement in a test and evaluation environment. Human Factors. 35(2), 263-281.

Isreal, J.B., Wickens, C.D., Chesney, G.L. & Donchin, E. (1980). The event-related brain potential as an index of display-monitoring workload. *Human Factors*, 22(2): 211-224.

Jones, D. & Farmer, E. (2001). *Applying the Cognitive Streaming Model to Air Traffic Management: A Preliminary Study*. QinetiQ Report QinetiQ/CHS/P&D/CR010575/1.0 (prepared for Eurocontrol, Brussels).

Jorna, P. (1997). *Human machine interfaces for ATM: objective and subjective measurements on human interactions with future flight deck and air traffic control systems*. Netherlands Aerospace Laboratory, Amsterdam.

Kirwan, B. & Ainsworth, L.K. (1992). *A Guide to Task Analysis*. London: Taylor and Francis

Kramer, A.F. (1991). Physiological metrics of mental workload: A review of recent progress. In D. Damos (Ed.), *Multiple Task Performance*. London: Taylor & Francis.

Lansdown T.C., Brook-Carter N. & Kersloot T. (2002). Primary Task Disruption from Multiple In-Vehicle Systems. *ITS Journal*, 7(2): 151-168.

Moray, N (Ed.) (1979). *Mental Workload: Its Theory and Measurement*. New York & London: Plenum Press.

Sternberg, S. (1969). Memory-scanning: mental processes revealed by reaction-time experiments. *American Scientist*, 57: 21-457.

Veltman, J.A., and Gaillard, A.W.K. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, 41 (5): 656-669.

Weston-Lovelock, K. & Abram, A. (1996). *Establishing the Reliability of Behaviourally Anchored Rating Scales at the RCB.* DERA, Farnborough, Report No. PLSD/CHS/HS3/CR96055/1.0.

Wierwille, W.W. & Connor, S., (1983). Evaluation of 20 workload measures using a psychomotor task in a moving base aircraft simulator. *Human Factors.* 25: 1-16.

Wierwille, W.W. & Eggemeier, F.T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors*, 35(2), 263-281.

Wilson, G.F. (2002). An analysis of mental workload in pilots during flight using multiple psycho-physiological measures*. The International Journal of Aviation Psychology*.

Wilson, G.F. & Eggemeier, F.T. (1991). Psychophysiological assessment of workload in multi-task environments. In D. Damos. *Multiple Task Performance.* London: Taylor & Francis.

Wilson, G.F. & Fullenkamp, P. (1991). A comparison of pilot and WSO workload during training missions using psychophysiological data. *Proceedings of the Western European Association for Aviation Psychology, Vol II, Stress and Error in Aviation,* 27-34, Brighton.

Wilson, G.F., Purvis, B., Skelly, J., Fullenkamp, P. & Davis, I. (1987). Physiological data used to measure pilot workload in actual flight and simulator conditions. *Proceedings of the Human Factors Society 31st Annual Meeting,* pp. 779-83, Santa Monica: Human Factors Society.

Yeh, Y-Y and Wickens, C.D. (1984). Why do performance and subjective workload measures dissociate? *Proceedings of the Human Factors Society 28th Annual Meeting*, 504-508.

Zeitlin, L.R. & Finkelman, J.M. (1975). Research note: Subsidiary task techniques of digit generation and digit recall as indirect measures of operator loading. *Human Factors*, 17:218-220

ABBREVIATIONS AND ACRONYMS

ATC	Air Traffic Control
ATM	Air Traffic Management
ß	In signal detection theory, a measure of the subject's response criterion in deciding whether a signal was present
BARS	behaviourally-anchored rating scales
d?	In signal detection theory, a measure of the discriminability of a signal from noise
DRAWS	Defence Research Agency Workload Scales
EEG	Electroencephalogram
EOG	Electro-oculogram
HR	heart rate
HRV	heart rate variability
ISA	Instantaneous Self-Assessment (of workload)
MCH	Modified Cooper-Harper
msec	millisecond
NASA	National Aeronautics & Space Administration
PDT	Peripheral Detection Task
RMS	root mean square
RT	Reaction time
SDT	signal detection theory
sec	second
SWAT	Subjective Workload Assessment Technique
TLX	Task Load Index, a set of workload scales developed by NASA