BEYOND BEDFORD; DEVELOPMENT OF A TWO-DIMENSIONAL PILOT WORKLOAD RATING SCALE

Garnet R Ridgway, Hannah R Shattock, Adam CR Hare
QinetiQ, MOD Boscombe Down, Salisbury, UK

Keywords: *Workload, Human Factors, Flight Test, Bedford*

Abstract

This paper presents the rationale for and development of a novel pilot workload rating scale, which seeks to address shortfalls identified in existing methods. The scale has been subjected to preliminary validation through both simulator assessments and rotary-wing flight test activities. The results show that the concept can be used effectively and has a number of advantages over existing methods.

1 Introduction

Accurate assessment of pilot workload is a key element of flight test, and one which has become significantly more important in the era of digital cockpits and increasing system automation. QinetiQ routinely undertakes such assessments in support of safe operations of the UK military aircraft fleet and on behalf of civil aircraft operators worldwide. Over the course of these assessments, a number of shortfalls have been identified with existing methods for quantifying pilot workload, most notably the Bedford workload rating scale [1]. This paper therefore outlines the development and validation of a novel workload rating scale, which seeks to address these shortfalls. The inception of this concept and supporting test activities were primarily conducted at MOD Boscombe Down in Wiltshire, UK; the proposed rating scale therefore bears that name.

2 Review of Existing Methods

A comprehensive review of all existing workload assessment methods falls beyond the scope of this Paper. However, the industry-standard tool is the venerable Bedford workload rating scale [1], which was itself a product of a QinetiQ precursor organisation. Whilst this has proven to be a valuable and valid tool, the following shortfalls have been identified during its use in aircraft assessments:

- **Time independence:** The two most important facets of aircrew workload are the maximum effort and the time for which that effort was applied. The Bedford scale provides only the former.
- **Subtlety of terminology:** The wording of the Bedford scale, whilst precise, requires significant scrutiny at the time of awarding ratings. This is not always feasible in a flight test scenario, and can be challenging for users whose first language is not English.
- **Reliance on supporting comments:** In isolation, a Bedford workload rating is not sufficient to draw conclusions about the nature of the workload encountered; heavy reliance is therefore placed on supporting pilot comments, adding to the time required to award ratings and to the scope for inaccuracies to encroach.

The above considerations have been addressed to varying degrees by other workload quantification methods such as the NASA TLX [2] and Instantaneous Self-Assessment (ISA) [3] scales. However, it is felt that no single existing scale provides a comprehensive workload assessment tool for this specific application.

3 Initial Conceptual Design

Based on the observed shortfalls in the existing methods, the requirements for the new scale were selected as follows:

- **Assessment of maximum applied effort:** The primary metric of pilot workload is an assessment of applied effort; this is most reliably given in terms of available capacity.
- **Application time of maximum effort:** The time for which the maximum effort is applied has a very significant effect on the safety impact of any areas of high workload.
- **Nature of workload drivers:** Workload is not necessarily cumulative across sub-tasks using different sensory channels. This can therefore aid with diagnosis of issues such as fixation or task shedding.
- **Unambiguous performance criteria:** The significance of an awarded rating should be immediately obvious, without the need for extensive interpretation.

It was intended that a rating scale based upon the above requirements would provide characterization of not only the amplitude of task workload, but also of its *shape*; i.e. the ability to differentiate between a short workload spike within a longer task and a moderate but relentless demand for the whole task duration. The degree of acceptability of workload would thence be determined based upon both a rating for maximum applied effort and application time. The initial concept for implementation is shown in Fig. 1.

A Boscombe Down Workload Rating Scale (BDWRS) rating consists of the three numbered elements shown in Fig. 1. For the maximum effort and application time parameters, a five point linear scale was selected based on successful use of this concept with the ISA scale. The third element consists of workload driver descriptors denoted by the addition of letters. These descriptors were selected on the basis of a review of commonly-reported workload drivers from assessment using the Bedford scale. For the preliminary scale, these were defined as follows:

- **Handling:** Physical interaction with primary flying controls. For example, compensation for aircraft handling deficiencies.
- **Communicating:** Communication either internally or externally. For example, internal crew communications.
- **Planning:** Any time a crew member is engaged in activities to support future actions. For example, inputting information into a navigation system.
- **Diagnosing:** Required when acquiring information about an unexpected occurrence. For example, finding root cause of an aircraft warning / caution / alert.
- **Monitoring:** Maintaining awareness of system parameters. For example, monitoring engine temperatures and pressures.
- **Other:** Any task not defined by the above workload descriptors.

The awarded BDWRS rating therefore takes the form of two numbers and a string of up to six letters, although it is not envisaged that more than three of the listed workload descriptors would be encountered in a single representative task.

The significance of the BDWRS rating can then be ascertained by plotting it on the axis shown in the lower portion of Fig. 1, allowing comparison against a set of performance criteria. The boundaries shown on the preliminary scale are based upon conceptual experience of pilot workload, e.g. that high effort can only be sustained for a short time, and vice versa.

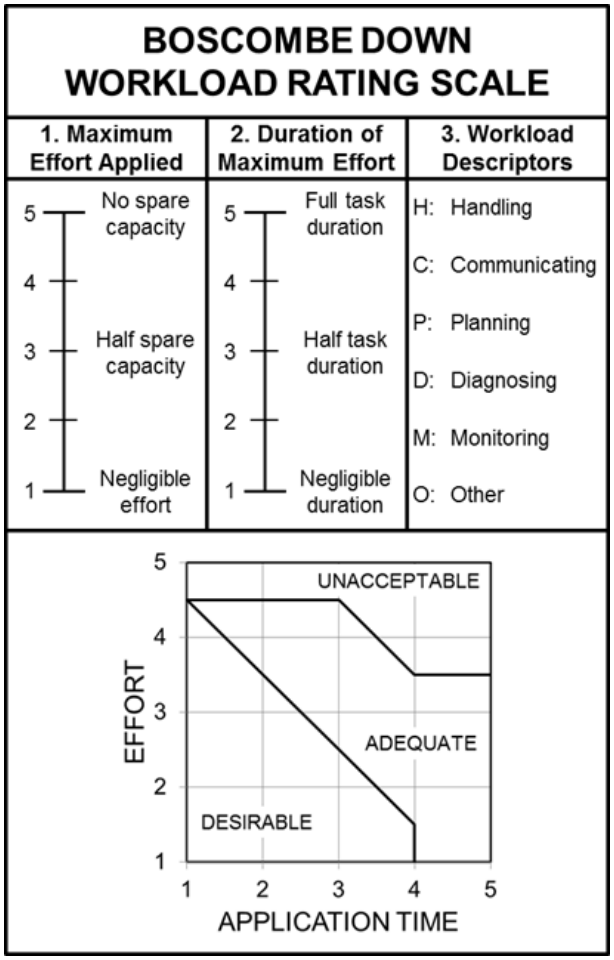

Fig. 1. Initial concept of the Boscombe Down Workload Rating Scale.

4 Overview of Evaluation Activities

The fundamental approach to the evaluation and validation of the BDWRS was to undertake role-representative flying tasks for which both Bedford and BDWRS ratings would be awarded. This would gather both general feedback and a body of evidence to allow refinement of the tool, particularly to the boundaries on the performance criteria plot. The overarching assumption applied to this process was that the output of the Bedford scale is a valid measure of aircrew workload, and that the three levels (1-3, 4-6 and 7-10) correspond to the "desirable", "adequate" and "unacceptable" criteria shown in Fig. 1.

All pilot input and ratings were provided by a pool of six qualified and current test pilots from the Rotary Wing Test & Evaluation Squadron at MOD Boscombe Down. The evaluation environments were as follows:

- **Boscombe Down Motion Simulator:** A fully configurable, six degrees-of-freedom motion simulator operated by QinetiQ for test aircrew training and research (Fig. 2). The aircraft model used for the purposes of this task was analogous to a Westland Lynx light utility helicopter.

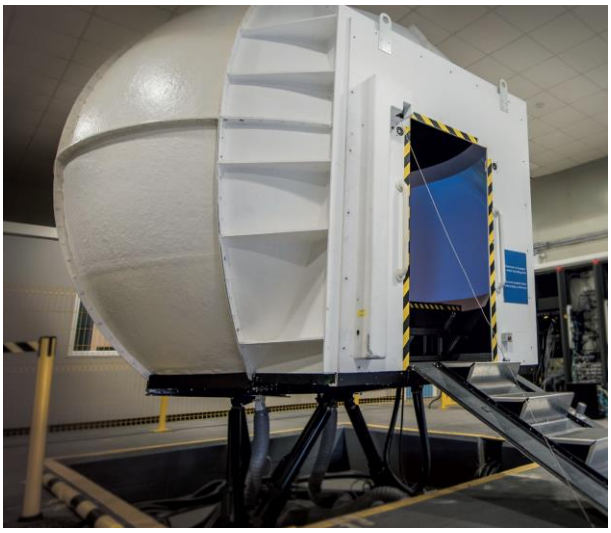

Fig. 2. Boscombe Down motion simulator.

- **Maritime Rotorcraft Simulator:** A high-fidelity, 6 degrees-of-freedom motion simulator used for aircrew training on a UK military maritime helicopter.
- **Light Utility Rotorcraft:** Two light utility helicopter types operated by QinetiQ for test pilot training / test and evaluation activities. An example is shown in Fig. 3.

Fig. 3. Light utility helicopter operated by QinetiQ.

The piloting tasks varied from procedural, non-handling tasks to dynamic role manoeuvres, with the aspiration of covering a wide range of task types and difficulty levels. These included (but were not limited to):

- Engine startup process;
- Hover taxy and standard Visual Flight Rules departure;
- Navigation system interaction;
- Interaction with cockpit systems whilst single pilot with varying levels of stability augmentation;
- Quick stops;
- Communication systems interaction;
- Deck landings;
- Instrument Flight Rules (IFR) approaches.

5 Preliminary Validation and Refinement

An initial proof-of-concept exercise was undertaken in the Boscombe Down Motion Simulator using the preliminary BDWRS (Fig. 1). The piloting task for this activity was selected on the basis of simplicity and repeatability. Specifically, pilots were instructed to perform a simple waypoint arrival task in which they were required to change heading upon arrival at a predetermined position. Additional elements were introduced in a pseudo-random sequence to vary task difficulty; for example, the requirement to speak to air traffic control, adjust transponder settings or to perform other routine cockpit interaction. A selection of the BDWRS and Bedford ratings awarded are shown in Fig. 4.

Task	Bedford	BDWRS		
		Effort	Time	Drivers
WP Arrival + Squawk	4	3	2	H
WP Arrival + Squawk + Report Speed	5	3	4	O
WP Arrival + Squawk (AFCS OFF)	8	4	3	OHM
WP Arrival + Squawk + Report Altitude (AFCS OFF)	7	3.5	3.5	HCO
WP Arrival + Squawk (AFCS OFF) + Radio Frequency change	9	5	5	HOM
Straight and Level + Radio Frequency change	5	3.5	2.5	O
Ground + Radio frequency change	7	2.5	4	O

Fig. 4. Example workload ratings for preliminary assessment.

Initial aircrew feedback was positive, suggesting that the concept was simple and intuitive to use. Additionally, the general agreement between low Bedford ratings and low BDWRS ratings suggests that the concept of using a two-dimensional scale for this purpose is fundamentally sound. Analysis of the full set of ratings and other feedback from aircrew yielded the following observations:

- Tolerance for operating with no spare capacity is higher than initially anticipated, providing it is only for a short time. This was supported by a number of ratings of 5, 1 and even 5, 2 being awarded low Bedford ratings. The common explanation for this was that many piloting tasks routinely require full attention, and that pilots are trained to expect / manage this for short durations.
- Pilots appear to be less tolerant of task demands which occupy a significant proportion of the task duration, even if the applied effort is relatively modest. This was attributed to the need to have unallocated time to maintain situational

awareness and to cope with emergent tasks.

- The "O: Other" workload driver descriptor was felt to be over-used and not sufficiently descriptive.

On the basis of the above, minor revisions were made to the BDWRS, specifically to the placement of the performance criteria boundaries and the replacement of "P: Planning" with "I: Interacting"; this sought to reduce reliance on "O: Other".

6 Further Validation and Refinement

Following completion of the initial proof-of-concept assessment, further data gathering activities were conducted using the aircraft and simulation devices described in Section 5. A total of 45 BDWRS ratings were recorded with corresponding Bedford rating in each instance. A small sample of the ratings is shown in Fig. 5.

Task	Bedford	BDWRS		
		Effort	Time	Drivers
Aircraft start: handling pilot	3	3	2	IM
Aircraft start: non-handling pilot	2	1	5	MC
VFR departure: handling pilot	4	3	3	H
VFR departure: non-handling pilot	4	3	4	CM
IFR recovery: AFCS holds engaged	3	2	4	CI
IFR recovery: AFCS holds disabled	5	3	5	H

Fig. 5. Example workload ratings for further assessment phase.

A number of points of particular interest were identified from the sample above. Firstly, for the aircraft start task, the Bedford ratings for the handling pilot and non-handling pilot correspond to "*Enough spare capacity for all desirable tasks"* and "*Workload low"* respectively. Without recording additional pilot comments, no further information is provided. By contrast the BDWRS ratings record that the handling pilot was applying a maximum of half of their workload capacity for less than half of the task duration, and that the main workload drivers were system interaction and monitoring. Whilst the non-handling pilot's overall level of workload was reported to be similar, the greater granularity provided by the BDWRS rating gives additional insight into the nature of the task; in this instance that a very low amount of effort was required throughout the entire task duration.

This concept is taken further for the Visual Flight Rules (VFR) departure task, for which both pilots awarded identical Bedford ratings (4: *"Insufficient spare capacity for easy attention to additional task"*). Without supporting comments, the demands placed upon the two pilots are apparently equivalent for this task. However, reference to the BDWRS ratings gives a more detailed and specific breakdown of their respective roles and the distribution of effort between them.

The IFR recovery task was repeated by the same pilot in the same aircraft but with Automatic Flight Control System (AFCS) modes enabled and disabled. Comparison of the associated BDWRS ratings shows the transformation of this task from one dominated by aircraft handling to a less demanding communicating and monitoring task. This provides an immediate, tangible and specific demonstration of the before / after benefits of using the aircraft's AFCS modes.

A primary aim of the validation activity was to refine the performance requirement boundaries on the BDWRS plot. This was achieved through a combination of the following:

- Comparison with Bedford ratings;
- Direct pilot feedback for each task;
- Engineering judgement based upon extensive experience of conducting aircraft workload assessments.

For the former of these, Bedford ratings were collated on a per-BDWRS rating basis. Specifically, for each point on the BDWRS effort / application time plot, the corresponding Bedford ratings were analysed in a number of ways. Most simply, the number of instances of particular Bedford ratings being awarded

against each BDWRS rating was plotted as shown in Fig. 6.

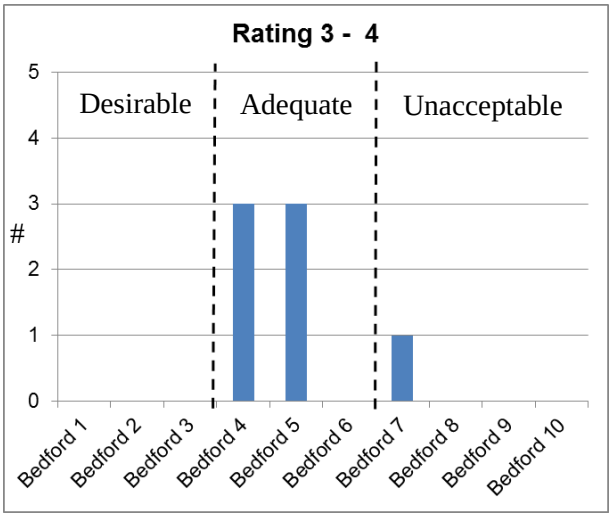

Fig. 6. Proportional breakdown of Bedford ratings awarded for BDWRS rating 3, 4.

This shows that the majority of Bedford ratings corresponding to BDWRS 3, 4 fall within the adequate band, suggesting that this point on the effort / application time plot should fall within the adequate category.

Conversely, Fig. 7 shows the same analysis for the BDWRS rating 3, 3. In this case, the ratings are evenly distributed astride the desirable / adequate boundary.

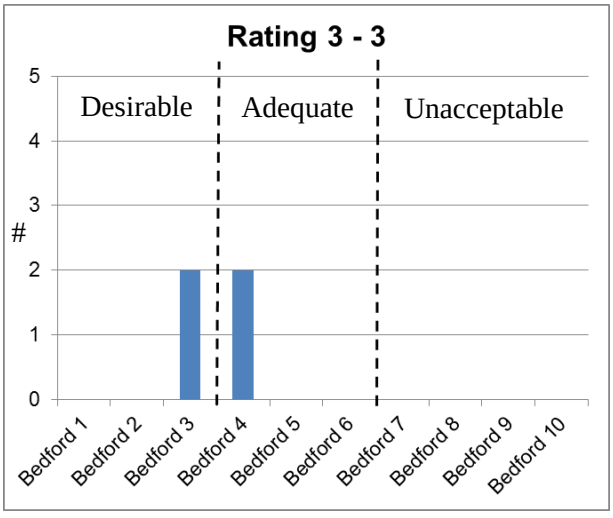

Fig. 7. Proportional breakdown of Bedford ratings awarded for BDWRS rating 3, 3.

This analysis was repeated for each BDWRS rating in order to inform performance boundary placement.

In addition to the above, overall trend analysis was also conducted, an example of which is shown in Fig. 8.

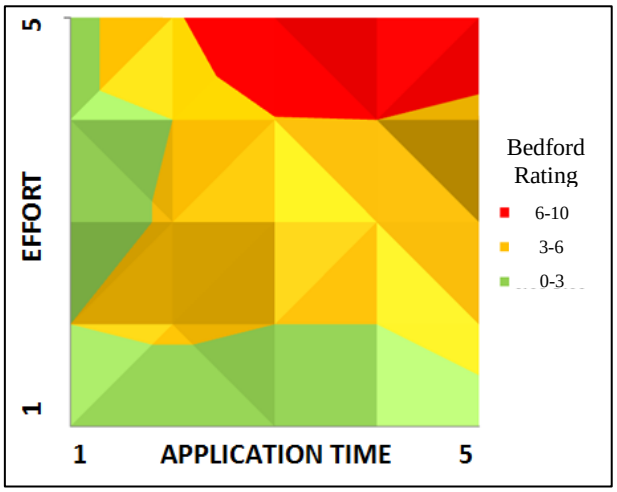

Fig. 8. Distribution of mean Bedford ratings across the BDWRS conceptual assessment space.

Fig. 8 shows the mean awarded Bedford ratings for each point on the effort / application time plot. Whilst it is acknowledged that mean Bedford ratings are in themselves not a valid measure of workload in a practical sense (Bedford ratings being non-linear and ordinal), their use in this way does provide a useful graphical representation of the trends involved. This serves two purposes: firstly, to verify the fundamental concept of the time / effort plot, i.e. that lower left corner is the region of lowest workload (green) and that upper right corner is the region of highest workload (red). Secondly, it augments the process for defining the performance criteria boundaries, the general trend for which can be seen in the three colour bands in Fig. 8.

The findings of each of the above validation activities were incorporated into a revised version of the BDWRS with supporting user instructions; this is shown in Fig. 9.

7 Considerations on General Usability

Throughout the course of representative use of the BDWRS, a number of pilot comments and third party observations were recorded. These are summarised as follows:

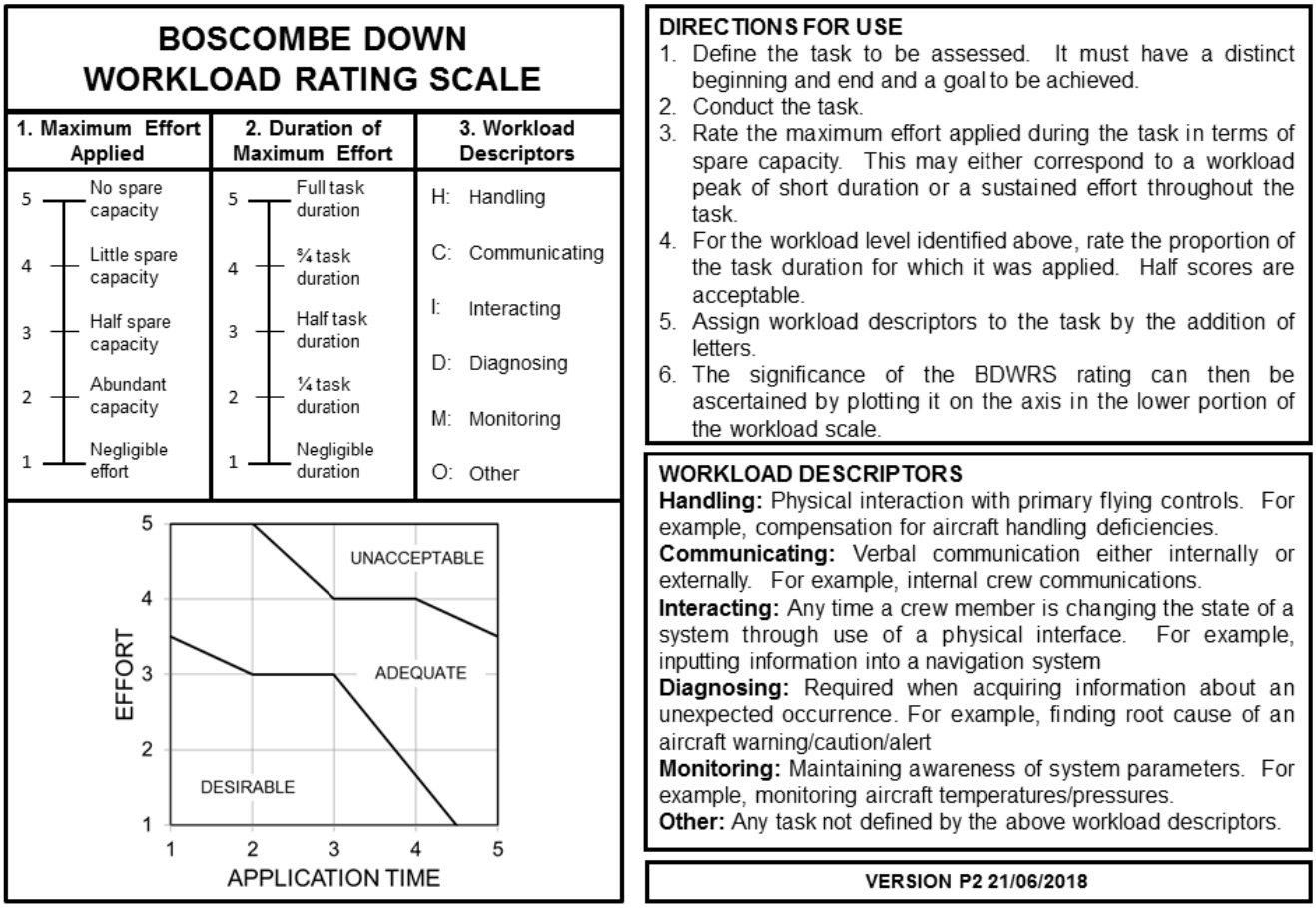

Fig. 9. The Boscombe Down Workload Rating Scale.

- Administering the BDWRS is significantly less time consuming than for the equivalent Bedford rating, even without supporting comments. For example, over a sample of 10 ratings, the average time for the BDWRS score to be awarded was 16 seconds; the corresponding figure for Bedford was 55 seconds.
- A potential source of confusion was identified with the orientation of the effort / application time plot axes. This is due to the effort score being awarded first, making it the intuitive choice for plotting on the horizontal axis. This can lead to errors in interpreting BDWRS ratings. Whilst this is acknowledged as a credible issue, it is felt that the benefits of retaining application time on the horizontal axis outweigh the minor potential source of confusion. In order to reduce the likelihood of this occurring, it is recommended that users are specifically briefed on this issue.
- The replacement of the subtle terminology of the Bedford scale with the more straightforward approach of the BDWRS was welcomed by users, particularly in the real aircraft environment.
- Similarly, the removal of a need to identify primary and secondary tasks prior to testing (a requirement for Bedford) was noted as a more efficient and less abstract approach.

8 Conclusions & Recommendations

This paper has summarised the initial design and validation of the BDWRS. The fundamental concept has been shown to be valid, and has a number of advantages over existing methods in terms of usability and precision. On the basis of the activities

conducted to-date, the authors make the following recommendations:

- Validity of the BDWRS could be further enhanced by additional data gathering activities. It should be noted that current validation has been limited to four rotary-wing types and a pool of six pilots; as such, data from other aircraft types would be particularly beneficial.
- This study has focused specifically upon aircrew workload as a result of the area of expertise of the authorship team. However, an advantage of the BDWRS is that it is context agnostic; as such, it could be applied to other areas such as land vehicles, general human-computer interaction or novel / semi-autonomous systems. Such an expansion of its usage would require appropriate validation.
- As with all tools which seek to quantify highly subjective entities, variation in results and repeatability are to be expected. Whilst not specific to the BDWRS, it is recommended that all users be appropriately briefed on the key concepts and that the scale is administered in accordance with the operating instructions.

Overall, the BDWRS concept is sufficiently mature to enable its use for aircraft workload assessments. To reflect the relatively small body of evidence supporting the performance criteria aspects (c.f. tools that have been in service for decades), BDWRS use should be supported by spot-checks against an industry standard method.

9 Acknowledgements

The authors would like to express their particular thanks to the Rotary Wing Test & Evaluation Squadron for their valuable input (and patience), without which this project would not have been possible.

This study was supported by the expertise, resources and facilities of QinetiQ.

References

[1] Roscoe, Ellis, *A subjective rating scale for assessing pilot workload in flight*, Royal Aircraft Establishment, 1990

[2] Hart, S G, *NASA-Task Load Index (NASA-TLX); 20 Years Later*, Human Factors and Ergonomics Society Annual Meeting. 50 (9): 904–908, 2006

[3] Jordan.C.S. *Experimental study of the effect of an instantaneous self-assessment workload recorder on task performance;* DRA Technical Memorandum (CAD5) 92011, 1992

Contact Author Email Address

For further information, please contact the lead author at grridgway@qinetiq.com

Copyright Statement