Development of Gauges for the QinetiQ Cognition Monitor

Andy Belyavin, Chris Ryder, and Blair Dickson

QinetiQ Ltd, Farnborough, UK
ajbelyavin@qinetiq.com

Abstract. This paper describes the development of a new version of the calibration procedure for the QinetiQ Cognition Monitor so that it can be implemented to support the development of a cognitive cockpit at NAVAIR. A new signal cleaning procedure for processing the electro-encephalogram (EEG) automatically is outlined and the results from tests in the UK and US are summarized. It is concluded that estimates of the content of the EEG signal at high frequencies are important to gauges measuring verbal and spatial workload. The combination of results from UK and US tests suggests that the cleaning procedure is effective, although increased robustness of the verbal gauge is desirable.

Keywords: EEG, signal cleaning, calibration, cognitive workload.

1 Introduction

At the heart of an adaptive Augmented Cognition system there has to be a methodology for the conversion of information that reflects the physiological or psychological state of the operator(s) into a series of gauges that relate to key aspects of human performance. In the majority of systems it has been assumed that the most important gauges should measure aspects of task demand (workload) in the context of human capacity restraints, and there is an implicit assumption that other forms of performance degradation will be captured. A prototype version of the Cognition Monitor that was designed to achieve these objectives was described in 2005 [1]. The development of the gauges for the prototype was based on three principles:

- Gauges should be based on a model of cognitive workload to provide face validity;
- Gauges should be calibrated on relatively pure tasks involving few mental processes since this provides a clear test of sensitivity;
- Amplitude-invariant measures should be used as input to the state classifiers as this supports reproducibility.

The Prediction of Operator Performance (POP) workload model [2] was used as the basis of the construction of the gauges, although the results were corrected for individual task performance. A monitoring task with varying levels of workload was used as the basis of calibration [1] and the measures of the Electro-encephalogram (EEG) were coherences and gains between electrodes in defined frequency bands.

D.D. Schmorrow, L.M. Reeves (Eds.): Augmented Cognition, HCII 2007, LNAI 4565, pp. 3–12, 2007.

This system was effective, but there were problems with porting it to a generic platform. So that the Cognition Monitor could be incorporated in a new version of the Cognitive Cockpit at NAVAIR the calibration procedure was redesigned to be executed reliably and rapidly with minimal operator intervention. In addition, the system was extended to include the development of distinct verbal and spatial gauges.

To meet the extended requirements of the Cognition Monitor a number of modifications to the overall approach were made. New calibration tasks were selected so that distinct verbal and spatial gauges could be calibrated. An automatic system for cleaning the data was developed and calculation of the gauge equations was standardized. The complete procedure was implemented in a single application to simplify execution. The new calibration tasks are described in Section 2. The cleaning procedure is outlined in Section 3 and the fitting procedure is described in Section 4, including the results from trials at QinetiQ and NAVAIR.

2 Calibration Tasks and Workload

The selection of tasks to serve as the basis of the calibration was a central component of the procedure. For the purposes of calibration it was decided that the tasks should satisfy three criteria:

- It should be possible to vary the task load in a controlled manner, so that a number of levels of workload could be defined;
- The task workload should be dominated by either verbal or spatial activity;
- The POP workload should be readily calculated.

There was substantial previous experience with two tasks that met all three criteria: the two-dimensional compensatory tracking task and the Bakan vigilance task [3]. Previous work indicated that compensatory tracking was dominated by spatial workload and that the Bakan task was dominated by verbal workload.

The spatial load was calibrated using the two-dimensional compensatory tracking task. The joystick provided first-order (velocity) control in both x and y directions and the disturbances were applied to the velocities of the cursor in both dimensions. The forcing function was constructed from separate combinations of six sinusoids for x and y, selected so that the pattern appeared random and the disturbance repeated only after a long interval. The frequency content of the forcing functions determines workload by affecting the frequency of operator interventions. This content was controlled through a single parameter, defined as *Dwork*, and the amplitudes were adjusted so that the root mean square error (RSME) of the cursor position in the absence of a participant response was independent of the value of *Dwork*, maintaining task difficulty.

A visual version of the Bakan vigilance task was used to generate verbal workload. Single digits in the range 0–9 were displayed in the middle of the screen for 500 milliseconds and successive stimuli were displayed at fixed inter-stimulus intervals. The target demanding participant response was a three-digit sequence of odd–even–odd digits. The rate at which targets appeared in the sequence was controlled at 5% of presentations. The Bakan task was displayed on the same screen as the compensatory tracking task, so that it was possible to choose either task alone or both together.

The cognitive workload associated with the compensatory tracking task was estimated using the POP model in the Integrated Performance Modelling Environment (IPME). The results were compared with experimental observations and a good match was found for *Dwork* set to 0.75 [2]. The model was used to estimate workload and RMSE for a range of values of *Dwork*, and the results for RMSE were compared with a small experimental test sample as displayed in Figure 1. There was good agreement between observed and predicted changes in RMSE and it was concluded that the estimates of workload for the tracking task could be used in the calibration.

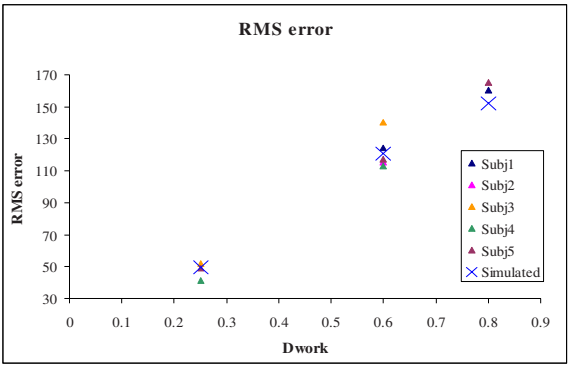

Fig. 1. Observed RMSE (Subjects 1–5) and predicted RMSE changes with *Dwork*

Previous work with the visual Bakan task established a simple relationship between the inter-stimulus interval and POP workload [2]. These results were incorporated in the IPME model so that estimates of workload for the dual-task case could be calculated and used in the calibration. The model indicated that there should be slight interference between the tracking and Bakan tasks, resulting in small increments in verbal or spatial workload in the dual-task case, and these were applied in the calibration procedure. The final version of the calibration procedure comprised 10 tests each of 150 seconds' duration:

- Single tracking task: *Dwork* = 0.25, 0.5 and 0.75;
- Single Bakan task: ISI = 0.75, 1.25, and 2.0 seconds;
- Simultaneous tracking/Bakan task: *Dwork* = 0.25/ISI = 2.0, *Dwork* = 0.25/ISI = 0.75, *Dwork* = 0.75/ISI = 2.0, and *Dwork* = 0.75/ISI = 0.75.

3 Pre-processing the EEG Data

The electro-encephalogram (EEG) was recorded from fourteen silver/silver chloride electrodes applied at the sites of the international 10-20 electrode placement system as indicated in Figure 2 [4]. All signals were amplified with respect to linked electrodes applied to the mastoid processes and vertical and horizontal derivations of the electro-oculogram (EOG) were recorded from electrodes placed above and below the right eye and lateral to the outer canthus of each eye respectively. A ground electrode was

applied to the mid-forehead. Data were digitized at a rate of 1024 samples/second, with a 500 Hz low pass filter and a high pass of 0.01Hz.

The recorded signals were analyzed in blocks of 4 seconds' length for tests using a single tracking task and blocks of length corresponding to the nearest multiple of the inter-stimulus interval for tests involving a Bakan task. The workload value assigned to a test is assumed to be uniform for the period and it is important to ensure that the activities within each analysis block are as uniform as possible.

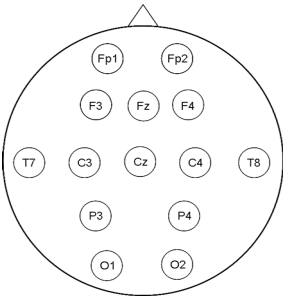

Fig. 2. Scalp placement of electrodes

For each block, multivariate kurtosis [5] was calculated using the data from the fourteen scalp electrodes. The contracted kurtosis tensor, C, was calculated where C is defined as

$$C_{ij} = k_{ijkl}\,\delta_{kl} \tag{1}$$

where summation is implied over repeated suffices, δ is the Kronecker delta and k_{ijkl} is the kurtosis tensor.

The eigenvalues and eigenvectors of C were calculated. If either the value of multivariate kurtosis or the largest eigenvalue of C did not support the rejection of the hypothesis of multivariate normality at 99%, no further analysis of the sample was conducted. If normality was rejected, a mixture of two normally distributed populations with a common zero mean but different variances was fitted to the one-dimensional projection of the sample onto the first eigenvector. Those observations for which the likelihood of belonging to the population with larger variance exceeded 0.5 were then corrected by projecting the sample value orthogonal to the first eigenvector. The procedure was then repeated for the second and third eigenvectors if the corresponding eigenvalues exceeded the 99% point for the largest eigenvalue. If any observations were corrected the complete procedure was repeated for the whole sample. If the sample was swept 12 times and there was still evidence for non-normality the block was rejected from the subsequent analysis.

The output from the pre-processing procedure is displayed in Figure 3. The dashed trace is the raw data and the superimposed solid trace is the "cleaned" data. From the data displayed in Figure 3 it is clear that the procedure correctly detects the effect of the blink visible on the EOG channel and provides a plausible correction to the observations. Apart from the major correction for the blink, the other adjustments in this block are relatively slight. For many blocks the adjustments include high-frequency corrections that are probably associated with temporalis muscle action. The procedure

was compared with direct regression on the EOG channels as a method of removing eye movement effects [6]. The kurtosis-based method did not prove completely effective at correcting for all eye movements, but it was found that the regression method occasionally *introduced* anomalies into the data and that on balance the current procedure is more satisfactory.

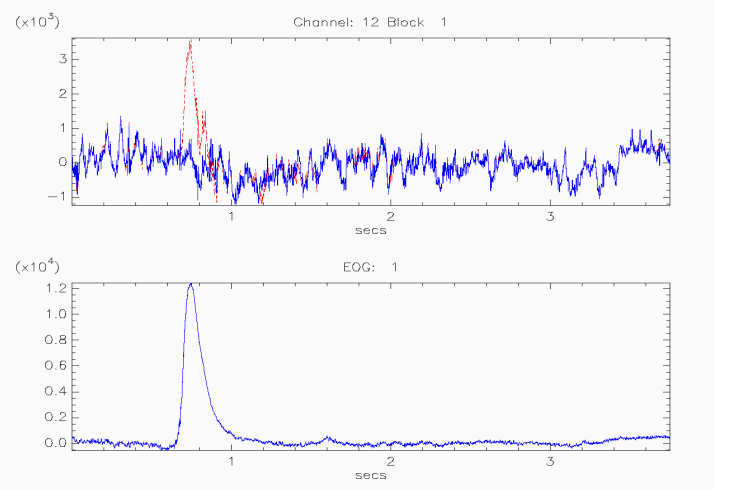

Fig. 3. Cleaned and raw data for a single 4 second block

Following data cleaning, the spectrum was calculated in each of the nine frequency bands defined in Table 1. The cross spectra between the 14 scalp electrodes in the nine bands were derived to yield estimates of the coherence between electrode pairs and the gains of one electrode relative to another, providing 1629 inter-correlated variables describing the EEG in each block.

Table 1. Frequency bands selected for spectral analysis

Band	Range UK analysis[1] (Hz)	Range US analysis[2] (Hz)
Delta	0.0 – 3.5	0.0 – 3.5
Theta	3.5 – 8.0	3.5 – 8.0
Alpha1	8.0 – 10.2	8.0 – 10.2
Alpha2	10.2 – 14.1	10.2 – 14.1
Beta1	14.1 – 20.0	14.1 – 20.0
Beta2	20.0 – 30.0	20.0 – 30.0
Gamma Low	30.0 – 47.0	30.0 – 47.0
Gamma Mid	53.0 – 70.0	47.0 – 57.0
Gamma High	70.0 – 100.0	63.0 – 100.0

[1] The bands exclude a small region around 50Hz.
[2] The bands exclude a small region around 60 Hz.These exclusions were made to disregard the predominant source of electromagnetic interference introduced from the mains power supply.

4 Calibration Procedure

4.1 Fitting the Calibration Equations

The prototype version of the Cognition Monitor used linear combinations of coherences and gains between electrode pairs to provide estimates of workload based on the EEG [1] and this approach was retained in the new version. Stepwise multiple linear regression was used to estimate the relationship between the estimates of workload in each condition and the EEG measures. The projected workload was treated as the dependent measure and the coherences and gains were treated as the independent measures, since the objective of the procedure was to derive the best estimates of workload conditional on the values derived from the EEG. A stepwise-up procedure was used to derive a preliminary estimate of the relationship including measures for which the t-value exceeded 3.29 and this was followed by a stepwise-down procedure in which measures were excluded if the associated t-value was less than 3.29. The quality of the overall fit was measured by the correlation coefficient, R^2, for the regression and the implied RMSE of the fit to workload.

The complete procedure, comprising selection of the calibration data files, data cleaning and calculation of the regression equations, has been implemented as two executable applications managed by a Java Applet. This Applet provides the user with a simple sequence of options to manage and monitor the calibration process and, more importantly, provides feedback on the quality of the calibration by presenting summary findings as displayed in the screenshot shown in Figure 4.

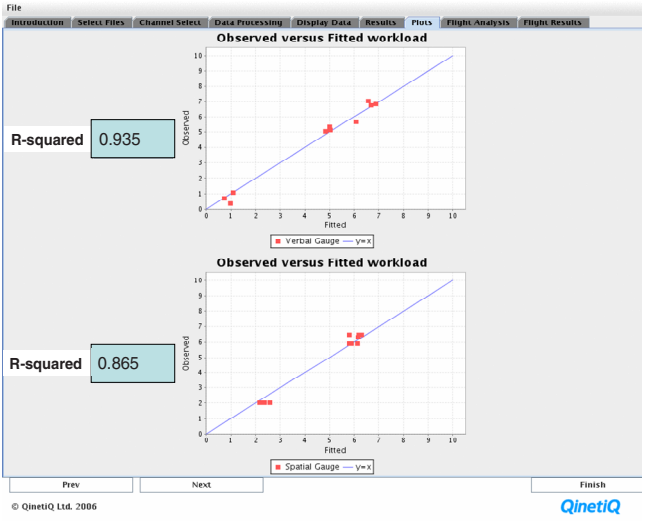

Fig. 4. Screenshot from the Java Applet showing the plots of observed and fitted workload

The points displayed in Figure 4 correspond to the means for observed and fitted values for each 150-second condition. The R^2 value *includes* the variation between 4-second blocks within the period in the values calculated from the EEG, since the

calibration is calculated at the level of the individual blocks. A high R^2 value implies both a good fit to the differences in mean values between the conditions and a relatively constant value within each condition, corresponding to good discrimination between workload levels.

4.2 Testing the Calibration Procedure

The complete calibration procedure was tested in the UK using a designed experiment. Eight participants were tested on two days. On both experimental days an extended calibration procedure of 15 test sessions was applied at the start of the experimental period, which included all the conditions of the standard calibration defined in Section 3, extended to cover all possible pairs of the six single-task conditions. After completion of the calibration, there was a short rest period, followed by three 5-minute tests in the Cognitive Cockpit with graded flying difficulty. Three test sessions of 150 seconds' duration with a different verbal logic task were recorded with controlled levels of inter-stimulus interval. The logic task involved reading statements concerning the status of a hypothetical vehicle system and then choosing an action that is consistent with the apparent state of the vehicle. Finally, the participant was tested for 10 minutes in the Cognitive Cockpit in which the logic task was executed for 30 seconds out of each minute.

The data were recorded for all sessions at a frequency of 1024 Hz. The calibration data were analyzed using the full procedure and estimates were made for Verbal and Spatial workload gauges. The data from the test sessions were cleaned using blocks with 4-second durations and the gauges applied to the cleaned data to derive estimates of the workload. The R^2 values were recorded for each calibration for Verbal and Spatial gauges. The estimates of the Verbal and Spatial gauges were collated for the flight and logic task sessions and were assessed for sensitivity using repeated-measures Analysis of Variance (ANOVA) treating Experimental Day (D), Task type (T), and Level of Workload (L) as fixed factors.

The final stage of the assessment of the overall procedure was conducted at NAVAIR. A single participant was tested using the basic calibration procedure described in Section 3 on 10 occasions. On a subset of the runs the participant was also tested in the cockpit and with the logic task. The R^2 values for the calibrations were recorded and, when available, the estimated workload gauges for the tests. No systematic analysis was conducted on the US results.

4.3 Results

The R^2 values generally exceeded 0.650 for both verbal and spatial calibrations in the UK and US, consistent with reasonable fits for the gauges. A plot of the values for the two gauges is displayed in Figure 5. As can be seen from the plot the majority of the values for spatial gauges are larger than the corresponding values for verbal gauges. In addition, the lower values mainly corresponded to earlier trials (Day 1 in the UK; earlier tests in the US).

The set of independent variables is clearly inter-correlated from the way it is constructed and there is considerable variation in the set of variables fitted in the calibration equations on each occasion. A count of the incidence of variables by electrode

and band was conducted for both verbal and spatial gauges and it was concluded that more than half the independent variables for the verbal gauge depended on measures in frequency bands 8 and 9 and more than a third for the spatial gauge were from the same source.

The analysis of the UK experiment indicated that there was a difference for both verbal and spatial gauges between the logic task and the flying task – $F = 7.32$, $df = 1,32$ $p < 0.05$ for the verbal gauge; $F = 10.76$, $df = 1,32$ $p < 0.01$ for the spatial gauge. In addition there was evidence for an effect of Day in the spatial gauge – $F = 9.30$, $df = 1,6$ $p < 0.05$. The means are displayed in Table 2. The residual error terms in both analyses indicate the level of variation for 4-second blocks in terms of their prediction – 1.821 for the verbal gauge and 0.911 for the spatial gauge.

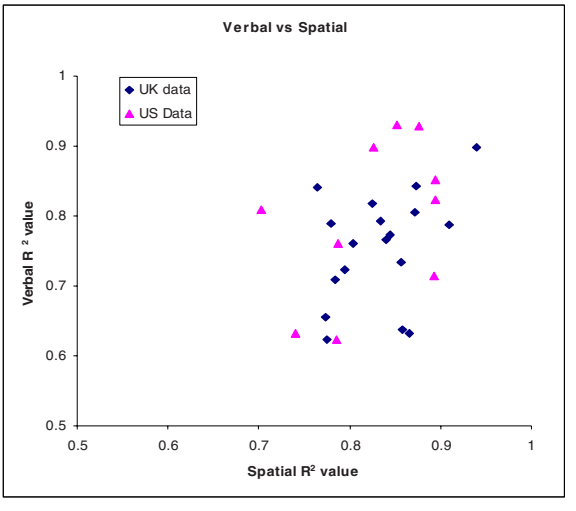

Fig. 5. Plot of the R^2 values for verbal gauges against spatial gauges for UK and US tests

Table 2. Means for the verbal and spatial gauges for the tasks in the UK experiment

		Verbal gauge		Spatial gauge	
		Day 1	Day 2	Day 1	Day 2
Low workload	Flying Task	3.18	3.88	6.24	4.92
	Logic Task	3.80	4.50	5.68	4.36
Medium workload	Flying Task	3.19	3.90	6.42	5.10
	Logic Task	3.81	4.52	5.86	4.54
High workload	Flying Task	3.43	4.13	6.40	5.07
	Logic Task	4.05	4.75	5.84	4.51

5 Discussion

For both the UK and US calibrations the EEG signal in the high frequency bands is an important component of both verbal and spatial gauges. There are problems in the use

of signals at these frequencies as they are vulnerable to contamination from electromagnetic interference from the mains supply, computer displays, EOG artefact (especially blinks) and temporalis muscle action. The cleaning procedure has been tested on signals that include known artefacts of this kind and it appeared to be effective at detecting and removing the majority of artefacts if applied to blocks of 4-second length.

Investigation of variation in the selected block length for cleaning and analysis indicated that the verbal gauge calibrated most successfully if the block length selected for cleaning and calibration was a multiple of the Bakan task inter-stimulus interval. Initial tests in the US employing a conventional analogue notch filter at 60 Hz to condition the signal prior to digitization almost completely eliminated any relationship between the EEG and verbal workload. On the basis of this evidence it is argued that the calibrations of verbal and spatial workload that use the high-frequency content in the EEG appear to be using content in the signal that relates to verbal or spatial activity. The main problem with the use of the high frequencies is the forced difference between UK and US analyses due to the different mains power supply frequencies. On the basis of the results obtained to date, it appears that the procedure works as well in the US as it does in the UK, despite the change in the banding of medium and high gamma activity.

The analysis of the UK designed experiment indicates that it is possible to distinguish the flying task from the logic task using the calibrated gauges. Analysis of individual runs in the US reveals that on most occasions the same result holds for similar patterns of flying task and logic task, although some calibrations of the verbal gauge appear to fail this test. The analysis of the UK data indicates substantially greater residual variability in estimates of verbal load than spatial load. It is not clear whether this pattern reflects true variability in verbal activity or failure of the calibration procedure. It is worthy of note that it is extremely difficult, perhaps impossible, to model very low verbal workloads due to the continual verbalisation that humans perform.

The revision of the procedure has provided a systematic and reproducible method for calibrating verbal and spatial workload gauges based on the EEG in a reasonable period of time. It is desirable that the procedure be extended to include other signals with different characteristics such as cortical blood oxygenation levels assessed using a Functional Near Infrared (fNIR) methodology to complement the EEG analysis and increase the robustness of the procedure. The use of high-frequency EEG signals is relatively novel and further work is needed to demonstrate that the overall procedure is robust.

References

1. Belyavin, A.J.: Construction of appropriate gauges for the control of Augmented Cognition systems. In: Proceedings of the 1st International Conference on Augmented Cognition, Las Vegas, NV (2005)
2. Belyavin, A.J., Farmer, EW.: Modeling the workload and performance of psychomotor tasks. In: 2006 Conference on Behavior Representation in Modeling and Simulation (BRIMS)-022. Baltimore (2006)

3. Bakan, P.: Extraversion–introversion and improvement in an auditory vigilance task. Br. J. Psychol. 50, 325–332 (1959)
4. Jasper, H.H.: The ten-twenty electrode system of the international federation. Electroencephalography and Clinical Neurophysiology 10, 371–375 (1958)
5. Mardia, K.V.: Measures of multivariate skewness and kurtosis with applications. Biometrika 57, 519–530 (1970)
6. Gratton, G., Coles, MG., Donchin, E.: A new method for off-line removal of ocular artifact. Electroencephalography and Clinical Neurophysiology 55(4), 468–484 (1983)