2005

Improving pilot mental workload classification through feature exploitation and combination: A feasibility study

Jeremy B. Noel
Air Force Institute of Technology

Kenneth W. Bauer, Jr.
Air Force Institute of Technology

Jeffrey W. Lanning
Air Force Institute of Technology

Noel, Jeremy B.; Bauer, Jr., Kenneth W.; and Lanning, Jeffrey W., "Improving pilot mental workload classification through feature exploitation and combination: A feasibility study" (2005). *U.S. Air Force Research*. 26.
http://digitalcommons.unl.edu/usafresearch/26

Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Computers & Operations Research 32 (2005) 2713–2730

computers & operations research

www.elsevier.com/locate/dsw

Improving pilot mental workload classification through feature exploitation and combination: a feasibility study

Jeremy B. Noel, Kenneth W. Bauer Jr.*, Jeffrey W. Lanning

Department of Operational Sciences, Air Force Institute of Technology, United States Air Force, AFIT/ENS, 2950 P Street, Wright-Patterson AFB, OH, 45433-7765, USA

Accepted 29 March 2004

Abstract

Predicting high pilot mental workload is important to the United States Air Force because lives and aircraft have been lost due to errors made during periods of flight associated with mental overload and task saturation. Current research efforts use psychophysiological measures such as electroencephalography (EEG), cardiac, ocular, and respiration measures in an attempt to identify and predict mental workload levels. Existing classification methods successfully classify pilot mental workload using flight data for a single pilot on a given day, but are unsuccessful across different pilots and/or days. We demonstrate a small subset of combined and calibrated psychophysiological features collected from a single pilot on a given day that accurately classifies mental workload for a separate pilot on a different day. We achieve classification accuracy (CA) improvements over previous classifiers exceeding 80% while using significantly fewer features and dramatically reducing the CA variance. Without the need for EEG data, our feature combination and calibration scheme also radically reduces the raw data collection requirements, making data collection immensely easier to manage and spectacularly reducing computational processing requirements.

Keywords: Artificial neural network; Backpropagation; Signal-to-noise ratio (SNR); Pilot workload; Mental workload; Feature exploitation; Feature identification and classification; Feature combination; Feature calibration; Feature space; Feature saliency

1. Introduction

Technological advancements in today's combat aircraft increase the demands on pilots, often requiring that their attention be split between multiple tasks. When divided attention is coupled with

* Corresponding author. Tel.: +1-937-255-6565x4328; fax: +1-937-656-4943.
E-mail addresses: Kenneth.Bauer@afit.edu, Kenneth.Bauer@afit.af.mic (K.W. Bauer).

0305-0548/$ - see front matter © 2004 Published by Elsevier Ltd.
doi:10.1016/j.cor.2004.03.022

stressful or mentally demanding situations, a potential for mental overload presents itself [1]. Studies of fighter aircraft pilots show how devastating the effects of mental overload can be. These pilots can become so involved in their current situation that they forget to perform critical tasks, such as G-force straining maneuvers. As a result, some pilots have lost consciousness and their lives. One pilot initiated a study regarding the problem after surviving a G-induced loss of consciousness (GLOC) incident [2]. He discovered that the United States Air Force lost 14 pilots due to GLOC over 10 years, with only one common factor found across the pilots: all but one of the fatalities occurred during mentally demanding portions of flight. Current research focuses on the idea that if a classifier can quickly and accurately analyze the psychophysiological data of a pilot and thereby provide insight into his current level of mental workload, then a system could be developed to reduce the possibility of future GLOC situations.

The Air Force Research Laboratory (AFRL)/Human Effectiveness Directorate (HE) at Wright–Patterson Air Force Base, Ohio, has conducted several studies on mental workload analysis in laboratory, simulator, and flight settings [3]. Their results indicate that the most influential psychophysiological features in classifying mental workload level are brain electrical activity, heart rate, breath rate, and eye blink measures [4–10]. Interestingly, however, research has shown that while feedforward multilayer perceptron neural networks show promising results classifying pilot mental workload from simulated flights using one set of psychophysiological features, a different set of features may be found to be most significant when classifying pilot data from actual flights [5,6,11].

Personnel working for AFRL/HE collected actual flight data using ten pilots each flying Wright–Patterson Aero Club Piper Cub aircraft on a specified route over 2 days. To collect the psychophysiological data, the pilots wore special recording equipment. Each flight produced large volumes of data that, when fully preprocessed, generates 151 features. Previous analysis of both simulator and flight data revealed that substantial feature reduction is attainable using a variety of statistical and analytical methods, with the signal-to-noise ratio feature-screening algorithm [12] producing the smallest feature set still capable of producing significant classification accuracy [6,11,13]. Furthermore, Laine et al. [11] and East et al. [13] found that artificial neural networks produce the most robust classifier for determining mental workload. They found that training an artificial neural network using reduced features sets over same-day, same-pilot data produced mental workload classification accuracies between 72% and 97%. However, the same-pilot over multiple days classifier yielded classification accuracy (CA) results around 50%, comparable to flipping a coin [13].

The focus of this effort is the development of a new feature combination and calibration scheme that exploits a small subset of psychophysiological features collected from a single pilot on a given day to accurately classify mental workload for a separate pilot on a different day. Extensive raw data preprocessing, including 29 Fast Fourier transformations for each second of flight data, prepared the feature data for analysis. The signal-to-noise ratio feature screening method is employed to determine the usefulness of 151 psychophysiological features in feed-forward artificial neural networks. Factor analysis is used to identify patterns in features that track associated changes in mental workload. Methodologies for workload level modification are tested to determine if they increase the accuracy of pilot mental workload measurement across pilots and days.

Exploratory factor analysis is used to show that the salient feature space varies by pilot and day. While artificial neural networks appear unable to fully discover this fact unaided, our new feature combination and calibration scheme appears to exploit a new feature space allowing us to more accurately discriminate between high and low mental workload. We demonstrate achieving classification accuracy (CA) improvements over previous classifiers exceeding 80% while using 97% fewer features and reducing the CA variance by over 95%. A considerable side benefit of our feature combination and calibration scheme is due to not requiring the use of EEG data, making data collection immensely easier to manage and dramatically reducing computational processing requirements. Along with the validated implementation method, the feature combination and calibration scheme appears to completely dominate all other classifiers over their entire operating curves and generally simplifies the entire classification process. The end result is that our feature combination and calibration scheme and its implementation method appear more practical than previous classifier and classification methods. Finally, the apparent identification of the new feature space also opens new doors for further improvements in classification accuracies.

The bottom line is that our feature combination and calibration scheme produces a single classifier from only one flight that appears able to more accurately predict pilot mental workload for other pilots and flights conducted on other days. These initial results open the possibility that the psychophysiological variations within and across individuals preventing previous methods from attaining acceptable classification accuracies may no longer present as major a hurdle.

2. Data collection

2.1. The experiment

The data used in our analysis are the same flight data described earlier. More specifically, ten volunteer pilots flew a predetermined flight route once a day for 2 days, accompanied by a technician from the flight propulsion laboratory and a copilot. The technician's job was to monitor the data collection process, and the copilot was present for safety reasons and was not part of the experiment. Each flight was divided into 22, 2-min flight segments. While ten pilots participated in the flight experiment, only the data from Pilots 1 and 4 were fully analyzed during the course of this research effort. Data from a third pilot (Pilot 6) became available later and was used for validation purposes.

The flight route was specifically designed to include three levels of mental workload: low, medium, and high. AFRL personnel estimated the difficulty of each flight segment before the flights were conducted, and the test pilots evaluated the difficulty of each flight segment after their flights. Fig. 1 shows a graph reflecting the pilot's subjective measures of workload associated with each flight segment. Understandably, there were some discrepancies between the researchers and the pilots concerning workload levels associated with each flight segment. For example, the pilots classified both the instrument flight rules (IFR) air work and visual flight rules (VFR) touch-and-go segments as high workload, while the researchers classified the VFR touch-and-go segment as high workload and the IFR air work as medium workload. Since both groups classified the touch-and-go segment of the flight as high workload, this flight segment became the minimum threshold for determining a high workload segment.

East et al. [13] found classifying three workload levels (low, medium, and high) very difficult and combined the low and medium levels into one group called low workload. This reduced the classification from a three-class to a two-class problem and also emphasized the primary objective of the research: accurately detecting high mental workload. Using the VFR touch-and-go flight segment as the threshold, the dark horizontal line in Fig. 1 identifies the split between the low (combined

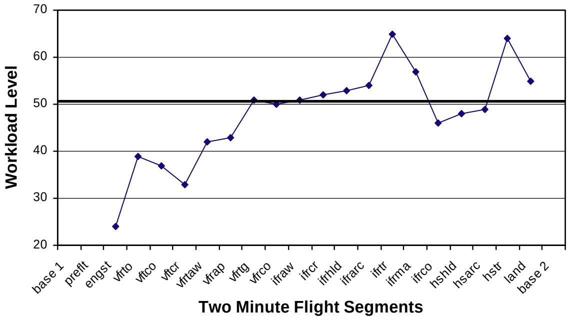

Fig. 1. Pilot subjective measure mental workload ratings.

with medium) and high workload levels. All flight segments below the line were defined as low mental workload and all flight segments above the line were defined as high mental workload. The creation of this line involves assumptions concerning workload level accuracy and flight segment transitions that could significantly increase classification errors.

The first assumption is that all flight segments were assigned the correct workload levels. It was assumed that all flight segments defined as low mental workload in fact represent equivalent workload levels. Similarly, it was assumed that all flight segments defined as high mental workload were of equivalent mental workload. Determining the true mental difficulty for individual flight segments is not a science and it is possible that the compromise between the researchers and pilots resulted in several inaccurate workload level definitions.

The second assumption is that the transition from low to high workload (or high to low workload) is instantaneous. In other words, at the workload transition point, the last second of one flight segment is correctly defined as low, and the first second of the next flight segment is correctly defined as high. However, transitions between mental workload levels are not really instantaneous since they occur over time and can vary by pilot.

While identification of these assumptions represents potential limitations to earlier efforts, our research using various schemes for defining different workload states found no apparent impact to CA [14].

2.2. Data collected

Four different types of psychophysiological data were collected during each flight: electroencephalography (EEG) data, ocular data, respiratory data, and cardiac data. The EEG data were collected at 256 Hz through 29 electrodes placed in a special cap worn by the pilots. The ocular, respiratory, and cardiac data were recorded in data files that contain the elapsed time in milliseconds between events. An event was the blink of an eye, the taking of a breath, or a beat of the heart. In order to make the data useful for analysis, the raw data were preprocessed. The same data preprocessing methods briefly addressed below were developed and used by Greene [6–8] and East [13].

The raw EEG data were collected and immediately sent through a program called *Manscan* 4.0, which filtered out some of the undesirable artifacts from the EEG signals such as muscle and eye movements. To remove the time dependency of the EEG signal, the raw data were passed through a Fast Fourier Transform (FFT). The FFT moved the data from the time domain into the frequency domain, which allowed estimates of power to be computed [15]. Five frequency bands were then filtered out of the EEG data: delta (1–3 Hz), theta (4–7 Hz), alpha (8–12 Hz), beta (13–30 Hz), and ultrabeta (31–42 Hz). Frequencies below 1 Hz or above 42 Hz are not associated with mental workload, so these data were not kept [16]. The power readings produced by the FFT for each of the five frequency bands were then summed to produce a total power reading for each band for that 1 s of data. The power was then averaged over 10 s. Five seconds of overlap was included in the calculation in order to smooth out the power readings from each electrode, resulting in 23 10-s average power readings for each electrode frequency band per 2-min segment. In summary, the EEG data produced 145 of the 151 total features for use when classifying the pilot's mental workload state, since each of the 29 electrodes produced a reading for the five frequency bands.

The preprocessing required for the remaining six physiological features comes from the heart, eye, and respiratory data files. Fortunately, preprocessing these data was less involved than the EEG data. Each of the heart, eye, and respiratory files produced two different features. In the case of the cardiac files, the two features were the heart rate (in beats per minute) and the heart rate variability. The heart rate variability is most easily thought of as the rate of increase or decrease in the heart rate over a period of time, which in this case was every 10 s. To preprocess the heart rate feature, the average beats per minute had to be computed. Since the data reflected the time between heartbeats (in milliseconds), the average time between beats for each 10-s window was calculated, and then inverted. After multiplying this result by 60, 000 ms per minute, the average beats per minute for each 10-s window was obtained. To calculate the heart rate variability feature, a first-order polynomial fit using ordinary least squares to the time intervals between heartbeats in each 10-s time window was completed. Next, the absolute value of the slope from the polynomial fit was retained to estimate the change in heart rate. The magnitude of this slope was used as the measure of heart rate variability.

The ocular and respiratory features were preprocessed in an identical manner to one another. To preprocess the number of blinks (or breaths) feature, the number of blinks (or breaths) that fell into each 10-s time window was counted. Fractional blinks (or breaths) were not considered, as they would naturally fall into a future 10-s time window. The preprocessing of the average time between blinks (or breaths) feature involved evaluating three different scenarios. If multiple blinks (or breaths) fell into a 10-s time window, then the simple average of the time between these blinks (or breaths) was used. On the other hand, if only one blink (or breath) fell in a 10-s time window, then the time between the last blink (or breath) in the previous window and the blink (or breath) in the current window was used. Finally, if no blinks (or breaths) fell into a 10-s time window, then the average time between blinks (or breaths) was determined by subtracting the time of the last blink (or breath) from the end of the current time window.

After preprocessing, the six physiological features were: heart rate (in beats per minute), heart rate variability, number of blinks (per 10-s time interval), interblink interval (time between blinks), number of breaths (per 10-s time interval), and interbreath interval (time between breaths). To allow EEG and physiological features to be included together within data sets, the same overlapping 10-s window method was employed. Combining the physiological, cardiac, ocular, and respiratory features brought the total to 151 features available for classifying mental workload.

One problem often encountered when using data from real test subjects versus simulated data is the possibility of having holes or gaps in the data. The data for this experiment had several cases where EEG features were missing for various lengths of time. Most likely, this was the result of a loss of contact between the pilot and one of the 29 electrodes. The options available to solve this problem include deleting each feature containing a gap from the data set, or filling the gap with non-zero data. If the first option is chosen and the entire feature is deleted from the data set, fair comparisons of feature sets for different pilots or for different days would require that the feature be removed from every data set. Should this feature be highly significant in predicting mental workload, then its removal could seriously affect the final selection of the most salient features and possibly the ANN's ability to accurately classify mental workload. If the gap is filled with non-zero data, then a decision must be made concerning how to best accomplish this action without losing the data integrity of the affected features.

The second option seemed most appropriate. We decided to keep the affected EEG features with missing data, and fill the gaps with average values based on the location of the gap. If the gap occurred in the middle of the data set, then the two data points immediately above and below the gap were used to create average values for filling the gap. If the gap occurred at the end of the data set, then the four data points immediately above the gap were used to create the average values for filling the gap. If the gap occurred at the beginning of the data set, then the four data points immediately following the gap were used to create the average values for filling the gap. The most likely effect of this procedure was a slight overall reduction in the total variance observed in each affected feature. We felt that accepting this slight reduction in variance was preferable to the total loss of the feature from the data sets.

3. Methodology

3.1. Feature selection

Artificial neural networks (ANNs) were chosen as the classification technique for this research effort. This decision was driven by the previous research results from East et al. [13] on these data, suggesting that ANNs produce more robust classifiers than other classification techniques. They tested quadratic and linear classification methods along with multilayer perceptron neural networks and identified classification accuracies ranging from between 72% and about 97% for a single pilot and flight depending on the method of feature selection and technique used. They found ANNs always either tied for or exceeded the highest classification accuracy of the other methods, and overall ANNs produced averaged classification accuracies several percentage points higher than the other methods. Using this information as background, this effort focused on ANNs created using the *Statistical Neural Network Analysis Package* (*SNNAP*) *Version 2.0* with an input layer, a hidden node layer, and an output layer. The number of nodes in the input layer had a one-to-one correspondence with the number of input features, and there were always two output nodes in the output layer signifying the two classes of mental workload. *SNNAP* produced a suggested number of hidden layer nodes, usually resulting in hidden layer nodes of approximately four times the number of input features.

Backpropagation was used as the training algorithm, and all activation functions were sigmoidal. Prior to training the ANNs, *SNNAP* standardizes the data sets to a mean of zero and a variance of one [17]. The data columns containing the known group memberships were not standardized, and remained 0 for low/medium mental workload levels and 1 for high workload levels. The training parameters for the ANNs included random initial weights between -0.1 and 0.1, the training rate set at 0.01, the momentum term set at 0.9, and the training termination rule of minimum training-test sum of square error.

The signal-to-noise ratio (SNR) saliency screening method [12] was used to reduce the 151 total available features to a smaller subset for classification. Previous feature reduction efforts on these data revealed that the SNR screening method developed a smaller set of features than other methods such as the SAS STEPDISK deterministic procedure [13]. Eq. (1) shows how the SNR method uses a direct comparison of a feature to an injected noise feature,

$$SNR_i = 10 \log \frac{\sum_{j=1}^{J} (w_{ij}^1)^2}{\sum_{j=1}^{J} (w_{Nj}^1)^2}, \tag{1}$$

where SNR_i is the saliency metric for the ith feature; j is the number of hidden nodes; w_{Nj}^1 is the weight connecting the injected noise feature (which is uniform (0,1)) to the hidden node layer; and w_{ij}^1 is the weight connecting the input feature to the hidden node layer. Since the weights connected to the noise feature tend to be small relative to the weights connected to more salient features, non-salient features have small ratios compared to salient feature ratios. Reducing the number of features is simply a matter of eliminating those features with the smallest SNR until only the most salient features remain.

Since many of the 151 features in each data set, especially the EEG features, are highly correlated with one another, and partially due to the randomness of the neural network initial weight values, different features can be selected for removal from the same network when identically initialized and trained several times [13]. With the high correlation among the features, any difference in feature selection should have negligible impact on the classification accuracy of the network, and so resolving feature selection differences is unnecessary. Across the different data sets, the classification accuracy for several neural networks starts to drop significantly (one or more percent) when fewer than 36 features remain, prompting the decision to keep as salient no more than 36 features per data set.

Past feature reduction efforts on these data have found that the number of salient features necessary to obtain high inter-day classification accuracy for individual pilots range from 5 to over 59 [13]. Our results indicate Pilot 1 has 35 salient features on day 1 and 28 salient features on day 2, while Pilot 4 has 36 salient features on both days 1 and 2. Feature reduction efforts conducted on mixed day data sets for an individual pilot revealed a different set of salient features. While this is similar to “peeking” into the future since the second day of data is not available for use when building a classifier using only the first day of data alone, some insights can be gained by studying the results. After combining the two data sets for each pilot into single large data sets, each data set was randomly split into training and validation data sets using a 65/35 ratio. In this effort, the training-test data set always consists of holdout exemplars from the training data set.

The results of the SNR saliency screening on multiple day data sets revealed that fewer features are salient for classifying Pilot 4 than Pilot 1. Specifically, we found Pilot 1 has 36 salient features and Pilot 4 has only 6. Furthermore, the features found most salient across the multiple day data sets were often different from those found most salient on individual day data sets. The possible causes for these differences other than the randomness of the initial weights in neural networks include wide variation in psychophysiological measures across days. This variation can be a result of differences in stress levels, sleep patterns, and caffeine levels, among other causes. It is also possible that humans exhibit an array of different physiological responses to the stress of high workloads.

3.2. Factor analysis

Factor analysis is based on the idea that the set of all features is explained by a smaller set of underlying factors. In the case of classifying mental workload, even though there are 151 different features, there may be a relatively small number of factors that drive these variables. The way these features are split into the different factors is derived from the covariances between the features. Factor analysis assumes that some of the feature variance is due to a common variance due to the factors, and the remainder is uniquely tied to the specific feature [18]. By performing factor analysis, the researcher hopes to identify and interpret the underlying factors to provide greater insight into the problem.

To perform factor analysis, the salient features in each data set were placed into the statistical software program *SAS JMP*. A separate scree plot was then built in Microsoft Excel using the eigenvalues from the covariance matrix of each data set. A scree plot is a plot of the ordered eigenvalues. The scree line helps determine how many eigenvalues to keep by establishing the number of factors to rotate using the varimax procedure in *SAS JMP*. The output of the varimax procedure is a factor loadings matrix, and this matrix was used to determine the psychophysiological feature-to-factor assignments. This was accomplished by assigning each feature to the factor with the largest absolute value factor loading for that particular feature. A review of the eigenvalues across several of the data sets revealed that the first eigenvalue represented approximately 15% of the total variation in the features, and the other eigenvalues each explained only 3–4% of the remaining variation. In order to capture a high degree of the total feature variation in these data sets, a large number of factors had to be kept. Keeping too many factors does not help reduce the dimensionality of the problem, and therefore limits the effectiveness of performing factor analysis. Keeping too few factors results in low factor loadings matrix values, making it difficult to determine which variables are really correlated to which factor, and also leads to difficulties with factor interpretation. By deciding to set the maximum number of factors to 20, sufficiently high factor loadings matrix values were produced, and it allowed for some useful groupings of features within and across the factors.

The decision to limit the number of factors to 20 enabled some interpretation of the factors, and more importantly, it highlighted key features within each factor that could be explored as we looked for patterns to exploit. With the relatively large number of factors for each data set, most of the factors ended up being associated with only a few of the features. This made factor interpretation somewhat easier given that brain researchers have identified that certain areas of the brain are associated with certain functions. A factor with only one feature assigned to it can be interpreted as being related to the function associated with that feature. Factor interpretation at this level, however, did not appear to provide direct insight into the research problem, and so further exploratory factor analysis was performed.

3.3. Exploratory factor analysis

The exploratory factor analysis consisted of two activities. First, the feature-to-factor assignments were compiled across the various data sets to find patterns among the factors. The feature-to-factor assignments were created by assigning each feature to the factor with the largest absolute value factor loading for that particular feature. Three compilation methods were used. The first compilation method involved grouping all of the feature-to-factor assignments by individual feature. The second

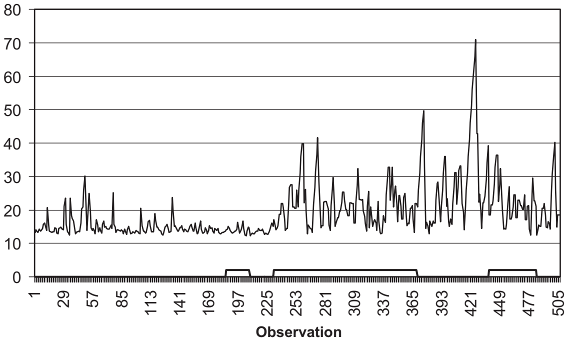

Fig. 2. Interblink feature for Pilot 1 on day 1.

method grouped the assignments by EEG node, which meant aggregating the five frequencies associated with each EEG node. The third method grouped the assignments by frequency, which meant aggregating across EEG nodes. These various compilations revealed that while the EEG features are evenly spread over all the retained factors, the physiological features are grouped rather tightly in the first six factors across the different data sets. In particular, the second factor showed a high concentration of the physiological features, with the ocular and heart features dominating the factor. As we will see, this observation played heavily in subsequent analyses.

The second exploratory factor analysis activity involved graphing the key feature-to-factor assignments in order to identify if any consistent patterns appeared within the data. A graph was made for each feature-to-factor association within the various data sets, representing the most important features across the factors. Most graphs revealed no discernible patterns across the mental workload levels; however, a few did show some interesting patterns. An example of a graph that showed a distinct pattern was the interblink feature from Pilot 1 on Day 1, shown in Fig. 2. The raised solid line along the bottom of the graph indicates periods of high mental workload. We noticed a definite increase in the value and variation of the interblink feature as the mental workload level increased from low to high.

Noticeable patterns for Pilot 1 were found only in the ocular features. We noticed a definite increase in the value and variation of the interblink feature as the mental workload level increased from low to high. In the number of blinks feature, we noticed a decrease during periods of higher mental workload. These patterns, however, were not as dramatic as seen on day 1. For instance, the amount of variability in the interblink feature, while certainly higher during periods of greater mental workload, was definitely not as variable as seen on day 1. Perhaps this decrease in variability was due to the learning curve effect caused by the identical flight path and same mental demands being repeated on the second day of the experiment. The increased familiarity possibly allowed Pilot 1 on day 2 to lower the visual concentration requirements necessary to execute the same maneuvers performed on day 1. Other features for Pilot 1 varied over time and mental workload levels, but they did not vary consistently like the ocular features.

Noticeable patterns for Pilot 4 were found only in the cardiac features. Unlike Pilot 1, Pilot 4's heart beats-per-minute (BPM) feature rose during periods of higher workload and stayed at an overall increased level throughout the higher workload periods. Furthermore, there was a visible decrease in the heart variability feature. No noticeable patterns in any of the EEG or respiratory features were found in the pilot data sets.

The different patterns in the psychophysiological features for Pilots 1 and 4 show that the pilots react differently under high workload conditions. Both pilots had two features that revealed patterns that displayed changes relative to mental workload, but the features were different for each pilot. Furthermore, we noticed features not exhibiting patterns for one pilot while exhibiting patterns for the other pilot look like noise features.

3.4. Feature combination and calibration scheme development

The apparent patterns found in the mental workload data through the exploratory factor analysis suggested an interesting possibility. It had become apparent that pilots react differently to increased workloads and that this reaction could be reflected through fundamentally different features. A linear combination of features was proposed [19]. The intent was to combine features in such a way that the sum increases dramatically when approaching high mental workload and drops dramatically when approaching low mental workload. This allows the ANN to "see across" the differing salient feature spaces. Following this concept, the features that appear to decrease when mental workload increases were subtracted from the linear combination, and the features that appear to increase when mental workload increases were added to the linear combination. Eq. (2) shows the linear combination and calibration scheme using standardized data,

$$\text{Calibration\_1} = -\text{Heart\_Variability}_{\text{SD}} + \text{BPM}_{\text{SD}} - \text{Blinks}_{\text{SD}} + \text{Inter\_Blink}_{\text{SD}}, \tag{2}$$

where $\text{Heart\_Variability}_{\text{SD}}$ is the standardized heart variability feature value, BPM_{SD} is the standardized heart beats-per-minute feature value, $\text{Blinks}_{\text{SD}}$ is the standardized number of blinks feature value, and $\text{Inter\_Blink}_{\text{SD}}$ is the standardized interblink feature value. Standardizing the feature data to a mean of zero and a variance of one was necessary because the feature data contained various units and magnitudes.

Graphing the Calibration_1 variable for the different data sets revealed a large amount of variability in the linear combination at any given mental workload level. In order to smooth this variability, three moving averages of Calibration_1 were added to complete the new set of features in the feature combination and calibration scheme. The lengths of the moving averages were 30, 60, and 120 s, and were labeled Calibration_30, Calibration_60, and Calibration_120. With the addition of the moving averages, the four features that comprise the feature combination and calibration scheme are Calibration_1, Calibration_30, Calibration_60, and Calibration_120. These four features totally replaced all 151 natural features when training ANNs using the calibration scheme.

4. Results

4.1. Results using feature combination and calibration scheme

Two different types of performance measures were used to assess the effectiveness of our proposed feature combination and calibration scheme: average CAs and receiver operating characteristics (ROC) curves. Average CAs were useful for summarizing a network's performance with categori-

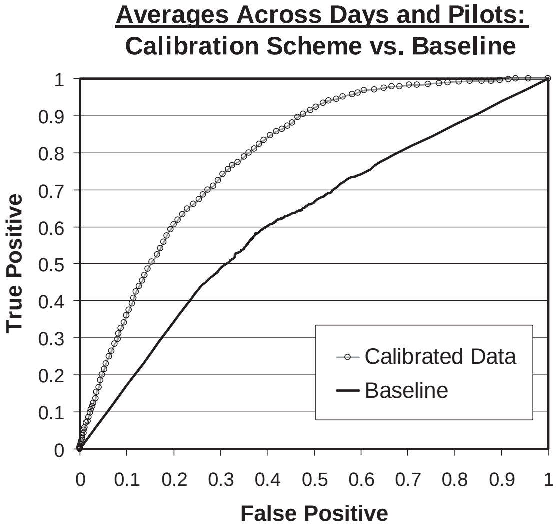

Fig. 3. Calibration scheme networks vs. non-calibrated baseline networks.

cal outputs in a single number; however, it implied equal costs of misclassification. In the case of determining pilot mental workload, we may be more interested in how accurately a network classifies high mental workload and less interested in how well it classifies low mental workload. The rationale here is that if we are interested in preventing GLOC situations, it is more important to correctly identify transitions to high mental workload than transitions to low mental workload. The ROC measure was especially useful when one category was more important than others, and provided network performance characteristics over a varying decision threshold. These two characteristics are probabilities of detection and false alarm, also known as true positive and false positive probabilities. For our application, the decision threshold represented the cut-off probability for detecting a signal and varied from 0.0 to 1.0.

Each average CA and ROC curve data point was based on 12 values, and never included the results from the same pilot and day combination used to train the network. For instance, if a network was trained using the data from Pilot 1 (day 1), a projection of this network would then be made using data sets from Pilot 1 (day 2), Pilot 4 (day 1), and Pilot 4 (day 2). No projection would be performed on Pilot 1 (day 1) since this represents the same pilot–day combination used to train the network. Another network would then be trained using data from Pilot 1 (day 2), and projections made for the three other pilot–day combinations: Pilot 1 (day 1), Pilot 4 (day 1), and Pilot 4 (day 2). This process would be repeated two more times using data from Pilot 4 (day 1) and Pilot 4 (day 2) to train the networks, and projecting the data sets from the other three pilot–day combinations through each network. We continued this leave-one-in fashion until we generated 12 projections, which when averaged together, become the average CA, or a single point on the ROC curve.

Fig. 3 shows the average result of networks trained using the feature combination and calibration scheme compared to the baseline. The baseline consists of networks trained using the 35 most salient features from each data set in addition to three moving averages per feature with lengths of 30, 60, and 120 s. As shown in the figure, the ROC curve developed using the feature combination

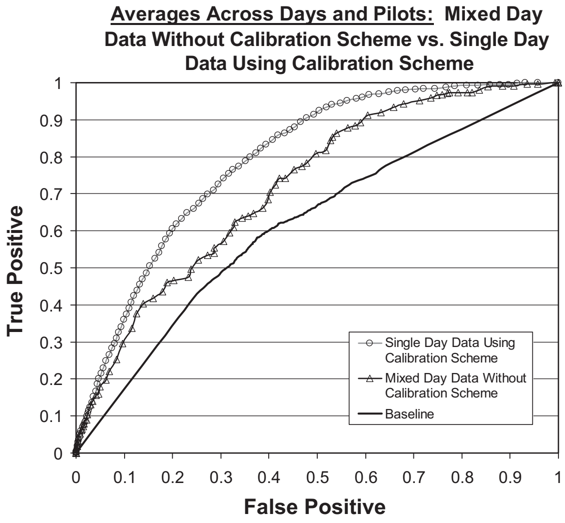

Fig. 4. Calibration scheme networks vs. non-calibrated mixed day network.

and calibration scheme completely dominates the baseline ROC curve. In addition, the average CA jumps from 60.11% to 72.02%, with individual calibrated network classifiers producing CA improvements up to 80% over comparable non-calibrated baseline network classifiers. Recall that previous researchers using this data obtained classification accuracies little better than 50% for the same-pilot over multiple days classifier [13].

Several modifications to the training data sets were made in an attempt to further improve network performance without success. The best results were consistently achieved using the combined feature and calibration scheme with all 22 flight segments in the training data sets.

Since our combined feature and calibration scheme makes use of only four ocular and cardiac features, another experiment was conducted where the ANNs were presented the same four ocular and cardiac features with information from all data sets mixed together. A random 60/40 split of the data built the training and validation data sets. Fig. 4 shows the results of this experiment. The average CA for the non-calibrated mixed day data ANN was 11.35% lower than the average CA for the calibrated full-day data ANNs (where all 22 flight segments were included in the training data sets). Fig. 4 shows that the combined feature and calibration scheme clearly improved network performance across the whole range of threshold values. Numerous increases in the number of hidden nodes in the hidden layer of the non-calibrated mixed day ANN did not improve either performance measure. These results indicate that the feature combination and calibration scheme provides additional information to the ANNs that they cannot produce themselves; as expected, the ANNs appear unable to identify the feature space found through the linear combination of the calibration scheme.

4.2. Results of validation effort

A validation effort was performed to fully determine the effectiveness and robustness of the feature combination and calibration scheme. The independent data set used for validation purposes came

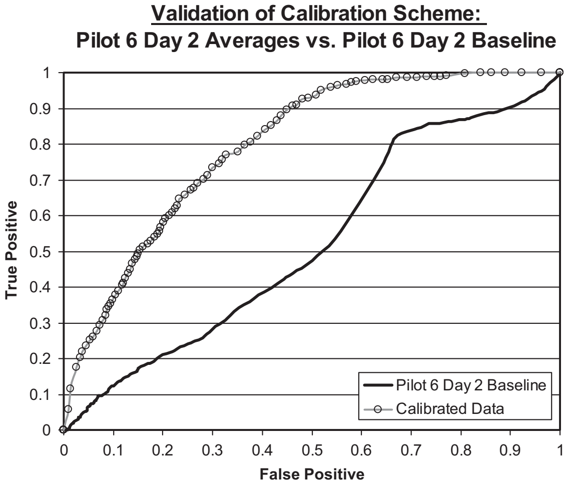

Fig. 5. Calibration scheme results compared to baseline results.

from Pilot 6 (day 2). To establish a baseline performance level, an ANN was trained using the most salient features in addition to the three moving averages per feature with lengths of 30, 60, and 120 s. After combining the features and calibrating the data following the feature combination and calibration scheme, another ANN was trained. The performance measures for the baseline and the feature combination and calibration networks were determined by averaging the results of four projections sent through the trained networks. The four data sets sent through the networks were: Pilot 1 (days 1 & 2), and Pilot 4 (days 1 & 2). Fig. 5 shows the ROC curve results. The average CA using the combination feature and calibration scheme jumped from 57.31% to 71.84%, an average improvement of over 25%. Furthermore, the ROC curve shows a large increase in true positive to false positive ratios across the whole curve. The performance measures in this validation effort indicate that the calibration method can be successfully applied to new data sets and potentially result in substantially improved pilot mental workload classification accuracy such as shown here.

An implementation methodology was developed and tested using the Pilot 6 (day 2) flight to see if the feature combination and calibration scheme could be implemented without knowing the true mean and standard deviation values for each of the four features included in the feature combination and calibration scheme. The implementation methodology was based on constantly computing throughout a flight the mean and standard deviation values for each of the four natural features used in the calibration scheme (heart BPM, heart variability, number of blinks, and interblink), and comparing these values to the minimum mean and standard deviation values identified after the 4-min point in the flight. Data from the first 4 min was used to baseline the implementation methodology because the first 4 min of each flight kept the pilots at low mental workload since they were only performing preflight checks with the engine off. The implementation methodology used only the larger of the minimum or actual values for standardizing the natural feature data and building the four combined and calibrated features.

The Feature Adjustment Factor Table, shown in Table 1 and used in the implementation process, is used when calculating the minimum mean and standard deviation values for each of the natural

Table 1
Feature adjustment factor table

Feature	Mean adjustment factor	Standard deviation adjustment factor
Heart variability	−0.3707	−0.2543
Heart BPM	0.2188	0.971
Number blinks	0.0115	0.0599
Interblink	0.1631	0.4328

features at the 4-min point in a flight. Each value in the table represents the average percent difference between the overall mean (or standard deviation (SD)) for a natural feature after a completed flight and the mean (or SD) for the same feature after only 4 min of flight. The table was constructed from four flights independent of the validation flight. To use the table and estimate the minimum mean and standard deviation values for natural feature i, we exercised Eqs. (3) and (4) shown below:

$$\text{Minimum mean}_i = (\text{mean}_i \text{ after 4 min}) \times (1 + \text{adjustment factor}_i), \tag{3}$$

$$\text{Minimum SD}_i = (\text{SD}_i \text{ after 4 min}) \times (1 + \text{adjustment factor}_i). \tag{4}$$

Since Table 1 reflects an average percent difference between the overall mean (or SD) for a natural feature after a completed 44-min flight and the mean (or SD) for the same feature after only 4 min of flight, the magnitudes and signs of the percent differences vary across the features. The signs of the percent differences are generally consistent with the trends observed in the feature means found during exploratory factor analysis, but the magnitudes of the percent differences are not particularly useful in determining the relative strength of the trends over time. It is important to note that the various feature trends identified during exploratory factor analysis were based on comparisons between periods of low and high mental workload while the plane was in the air. Since the engines were not started until after the 4-min point of each flight, the pilots remained in a lower than normal workload state when compared to the other periods of low mental workload during flight. This period of artificially low mental workload, plus the fact that there were more low workload than high workload periods during the flight, limits the value of comparing the feature averages and standard deviations at the 4-min point to the feature averages after the entire flight. What Table 1 does reflect is the average percent difference for the means and standard deviations of the features as the pilots transition from an exclusively very low mental workload state to an overall elevated mental workload state mixed with periods of high and low mental workload. For example, the mean adjustment factor for heart rate variability is −0.3707, consistent with the observation that heart rate variability tends to decrease during periods of increased mental workload. The relatively large magnitude of this adjustment factor indicates that the pilots remained in a much lower average mental workload state during the first 4 min of each flight than they averaged in the remaining periods of the flight. The sign and magnitude of the SD adjustment factor for heart rate variability, −0.2543, indicates less variability in the feature as time passed and the overall average mental workload increased. Similar observations can be made concerning the positive signs of the heart rate (BPM) and interblink mean adjustment factors (0.2188 and 0.1631, respectively). Both positive values parallel the observations that the feature means tend to increase during periods of increased mental workload, and their moderate magnitudes indicate that

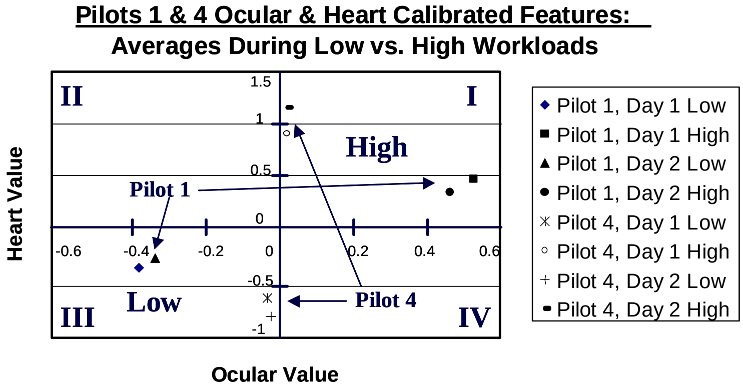

Fig. 6. Average calibrated feature values by mental workload level.

the pilots remained at lower average mental workload during the first 4 min of the flight than during the remaining flight periods. The relatively large SD adjustment factors for these features (0.971 for heart rate and 0.4328 for interblink) reflect increased feature variability as the flight progressed. The only feature whose mean adjustment factor did not mirror the observations made during exploratory factor analysis is the number of blinks feature. The small positive value of its mean adjustment factor (0.0115) is inconsistent with the general observation that the number of blinks tended to decrease during periods of increased mental workload. This inconsistency is likely due to the relatively small amount of increase or decrease observed in the ocular features (and the number of blinks feature in particular) when shifting between low and high mental workload compared to the changes observed in the cardiac features. Fig. 6, discussed in greater detail later, visually shows this observation. It highlights the average magnitude differences between the combined ocular and cardiac features when shifting between low and high mental workload for Pilots 1 and 4. Upon inspection, the relative magnitude difference observed as the mental workload shifts between low and high is much less with the combined ocular feature than the magnitude difference observed with the combined cardiac feature. In addition, since not all pilots experienced the same amount of change in each feature as mental workload varied, the small amount of change in the number of blinks feature observed in some pilots was likely overcome by the magnitude of variations observed across the features for the various pilots on the different flights. The relatively small SD adjustment factor for this feature (0.0599) further reinforces the notion that little average variation in the feature occurred across the pilots as mental workload levels changed.

Our implementation process began by assuming the pilot would remain in a low mental workload state during the first 4 min of flight while the preflight checks are performed. Consequently, we set the Calibration_1 feature to -1.0 during this period to signify low workload. Since the other three calibration features were moving averages of the Calibration_1 feature, they also had values of -1.0 during this 4-min period. After 4 min of flight, we used the Feature Adjustment Factor Table and Eqs. (3) and (4) to estimate the minimum mean and standard deviation values for the rest of the flight. As time passed and the four natural feature values became available, they were standardized based upon the larger of the minimum mean and standard deviation values or the actual mean and standard deviation values. The Calibration_1 feature was then computed using Eq. (2), and the other three moving averages combined and calibrated features were updated. The four combined and calibrated features were then presented to the trained ANN for a prediction of current mental workload, and this process was repeated until the end of the flight. If this approach was implemented operationally, then after each completed flight the Feature Adjustment Factor Table should be updated to reflect the new pilot information. Alternatively, a personalized Feature Adjustment Factor Table could be built using data exclusively from one pilot. Steps 1–5 summarize the steps in this implementation process.

1. For the first 4 min of flight, set the Calibration_1 feature to -1.0 to reflect the assumed low workload state of the pilot. After 4 min of flight, compute the actual mean and standard deviation for each of the four natural features used in the feature combination and calibration scheme.
2. Estimate the minimum mean and standard deviation for each natural feature using Eqs. (3) and (4). These minimum values are found by multiplying the actual mean and standard deviation values by the appropriate adjustment factor from the Feature Adjustment Factor Table, shown in Table 1.
3. As each set of four natural features becomes available during the flight, the continually updating mean and standard deviation for each natural feature is compared to the minimum values found in Step #2. If a natural feature mean or standard deviation value falls below the respective estimated minimum value, then the minimum value is substituted when standardizing the feature. If a natural feature mean or standard deviation rises above the respective minimum value, then the larger value is used when standardizing the feature.
4. Using the mean and standard deviation values from Step #3, compute the Calibration_1 combined calibration feature, and update the three moving average combined calibration features: Calibration_30, Calibration_60, and Calibration_120. Present the four combined calibration features to the trained ANN for a prediction of current pilot mental workload.
5. Repeat Steps #3 and #4 until the end of the flight. Before the next flight begins, update the Feature Adjustment Factor Table until the table values stabilize. Alternatively, update a personalized Feature Adjustment Factor Table for exclusive use by an individual pilot.

Results of the implementation experiment revealed an ROC curve nearly identical to the full feature combination and calibration ROC curve, and a decrease in average CA compared to the full feature combination and calibration results of only 2%. These performance measures provided preliminary indications that the implementation methodology was robust and accurately reproduced the full feature combination and calibration benefits.

5. Conclusions

The feature combination and calibration scheme presented in this paper, including the implementation methodology, appears feasible and produced superior results by finding a new feature space unable to be found by the ANNs themselves. Accurate mental workload classification requires finding the appropriate feature space for each individual, and we have shown that this feature space can vary by pilot and time. Fig. 6 visually shows the different feature spaces between Pilots 1 and 4 by comparing average combined and calibrated ocular and cardiac feature values during periods of low and high mental workload. Both pilots shift between quadrant III and quadrant I as their mental workload levels change between low and high; however, Fig. 6 clearly shows that their combined and calibrated feature values shift along different axes. As a result, networks trained using the non-calibrated feature data for a single pilot on a given day stand little chance of accurately classifying mental workload for a separate pilot, and the large psychophysiological differences observed for an individual pilot over time allow only a slightly better chance of accurately classifying mental workload for the same pilot on a different day.

The feature combination and calibration scheme appears to reduce the impacts of the psychophysiological variations that occur across different pilots and over different days. If one or more of the four features included in the feature combination and calibration scheme were not significant to a particular pilot on a certain day, then those features basically represented small amounts of noise. Their inclusion in the linear combination resulted in the addition of this noise. Before a network was trained, however, the neural network software standardizes the data, thus mitigating the effect of insignificant features. As a result, the linear combination calibration scheme allowed the significant features to provide valuable mental workload information to the network, and rendered the effects of the other features as insignificant. For example, consider a network trained for each pilot on either day. The feature combination and calibration scheme adds the normalized contributions from the interblink feature, subtracts the contribution from the blink feature, adds the contribution from the heart BPM feature, and subtracts the contribution from the heart variability feature. For Pilot 1, the heart variability and heart BPM features are insignificant so their additions to the combination and calibration scheme are really additions of noise. As mentioned before, Pilot 4 does not display the same consistent patterns as Pilot 1 in the ocular features, but Pilot 4 does have two consistent patterns in the heart BPM and heart variability features. This results in two features added to the combination and calibration scheme for each pilot that provide information about mental workload and two features that add noise. The outcome of the combination and calibration scheme is a new combined and calibrated feature for each pilot containing useful information about mental workload, which can be directly compared to the same new combined and calibrated feature developed for other pilots. This research suggests that it might be possible to overcome the large psychophysiological variations within and across pilots and presents a new feature combination and calibration scheme that may help overcome a long-standing stumbling block to achieving higher classification accuracy and good ROC curve performance.

Our research indicates that the feature combination and calibration scheme dramatically improves our ability to accurately predict pilot mental workload. Furthermore, our validation effort results suggest that the feature combination and calibration scheme is robust, and the implementation method results identify that the calibration scheme can be successfully implemented without any apparent significant loss of predictive capabilities.

Several opportunities exist for further research on feature combination and calibration to enhance the classification of pilot mental workload. The first opportunity involves exploring combination and calibration schemes other than the linear combination presented in this research. Examples include schemes containing interaction terms and non-linear functions. The second opportunity applies optimization techniques for improving the weighting of the features within the combination and calibration scheme to optimally highlight the changes in mental workload level. Provided the predictive power and operating characteristics of the combination and calibration scheme meets warfighter needs, the third opportunity includes moving the feature combination and calibration scheme and the implementation methodology towards additional testing and future system development.

Acknowledgements

We would like to thank the Air Force Research Laboratory Human Effectiveness Directorate (AFRL/HE) and the Air Force Office of Scientific Research for their support of our research.

References

[1] Hankins TC, Wilson GF. A comparison of heart rate, eye activity, EEG, and subjective measures of pilot mental workload during flight. Aviation, Space, and Environmental Medicine 1998;69:360–7.

[2] Auten J. G-LOC: is the cluebag half full or half empty? Flying Safety 1996;52:5–6.

[3] Air Force Research Laboratory, AFRL. Flight Psychophysiological Laboratory. Office Brochure, Flight Psychophysiological Laboratory, Human Interface Technology Branch, Crew System Interface Division, Human Effectiveness Directorate (AFRL/HECP); 1998.

[4] Wilson GF. Applied use of cardiac and respiration measures: practical considerations and precautions. Biological Psychology 1992;34:163–78.

[5] Wilson GF. Air-to-ground training missions: a psychophysiological workload analysis. Ergonomics 1993;36(9): 1071–87.

[6] Greene KA, Bauer KW, Wilson GF, Russell CA, Rogers SK, Kabrisky M. Selection of psychophysiological features for classifying air traffic controller workload in neural networks. International Journal of Smart Engineering System Design, 1998.

[7] Greene KA, Bauer KW, Kabrisky M, Rogers SK, Wilson GF. Estimating pilot workload using Elman recurrent neural networks: a preliminary investigation. In: Dagli CH, et al., editors. Intelligent Engineering Systems through Artificial Neural Networks, vol. 7. New York: ASME Press; November 1997a.

[8] Greene KA, Bauer KW, Kabrisky M, Rogers SK, Wilson GF. Estimating pilot workload using Elman recurrent neural networks: a preliminary investigation. In: Dagli CH, et al., editors. Intelligent Engineering Systems through Artificial Neural Networks, vol. 7. New York: ASME Press; November 1997b.

[9] Wilson GF, Fisher F. Cognitive task classification based upon topographical EEG data. Biological Psychology 1995;40:239–50.

[10] Wilson GF, Gundel A. Topographical changes in the ongoing EEG related to the difficulty of mental tasks. Brain Topography 1992;5:17–25.

[11] Laine TI, Bauer KW, Lanning JW. Multiple crewmember workload classification using neural networks with input feature selection. Intelligent engineering systems through artificial neural networks. Proceedings of Artificial Neural Networks in Engineering (ANNIE) International Conference, St. Louis MO, 7–10 Nov 1999.

[12] Bauer KW, Alsing SG, Greene KA. Feature Screening using Signal-to-Noise Ratios. Neurocomputing 1999;31: 29–44.

[13] East JA, Bauer KW, Lanning JW. Feature selection for predicting pilot mental workload: a feasibility study. International Journal of Smart Engineering System Design 2002;4:183–93.

[14] Noel JB. Pilot mental workload calibration. MS thesis, School of Engineering, Air Force Institute of Technology, Wright-Patterson AFB OH, March 2001.

[15] Math Works, Inc. MATLAB signal processing toolbox user's guide. Natick MA: Math Works, 1998. 2-2, 3-5–3-8.

[16] Wilson GF, Fullenkamp P, Davis I. Evoked potential, cardiac, blink, and respiration measures of pilot workload in air-to-ground missions. Aviation, Space and Environmental Medicine, February 1995; 100–5.

[17] Wiggins VL, Borden KM, Turner KL, Looper LT, Grobman JH. Statistical Neural Network Analysis Package (SNNAP) Version 2.0 User's Manual, Interim Technical Paper- July 1994–July 1995, Air Force Material Command, Brooks Air Force Base TX; 1996.

[18] Bauer KW. OPER 685, Applied Multivariate Data Analysis. School of Engineering, Air Force Institute of Technology, Wright-Patterson AFB OH; Fall 2000.

[19] Noel JB, Bauer KW, Lanning JW. Pilot mental workload calibration. Intelligent engineering systems through artificial neural networks. Proceedings of Artificial Neural Networks in Engineering (ANNIE) International Conference, St. Louis, MO, 4–7 Nov 2001.