Real-Time Workload Assessment as a Foundation for Human Performance Augmentation

Kevin Durkee[1], Alexandra Geyer[1], Scott Pappada[1], Andres Ortiz[1], and Scott Galster[2]

[1] Aptima, Inc., USA
[2] Air Force Research Laboratory, USA
{kdurkee,ageyer,spappada,aortiz}@aptima.com,
scott.galster@wpafb.af.mil

Abstract. While current military systems are functionally capable of adaptively aiding human operators, the effectiveness of this capability depends on the availability of timely, reliable assessments of operator states to determine when and how to augment effectively. This paper describes a response to the technical challenges associated with establishing a foundation for reliable and effective adaptive aiding technologies. The central component of this approach is a real-time, model-based classifier and predictor of operator state on a continuous high resolution (0-100) scale. Using operator workload as a test case, our approach incorporates novel methods of integrating physiological, behavioral, and contextual factors for added precision and reliability. Preliminary research conducted in the Air Force Multi Attribute Task Battery (AF_MATB) illustrates the added value of contextual and behavioral data for physiological-derived workload estimates, as well as promising trends in the classification accuracy of our approach as the basis for employing adaptive aiding strategies.

Keywords: Workload, Augmentation, Human Performance, Modeling and Simulation, Physiological Measurement.

1 Introduction

To address the modern threat environment, military operations must overcome a variety of demands and resource constraints, such as manpower limitations, information overload, sustained long-term missions, and an increasingly complex decision space. This reality leads to our military force being more vulnerable to performance decrements related to increases in cognitive workload, stress, and fatigue. There are available technological solutions that could help mitigate these types of performance decrements through adaptive aiding and, consequently, benefit the effectiveness of active operational systems. Traditional approaches to designing user interfaces (UI) typically result in a fixed presentation of information throughout the entirety of the operator interaction with a control station; however, human operator states (e.g., workload, engagement, and affect) are dynamic. For instance, if the system detects that the human operator is experiencing high workload, as when an remotely piloted aircraft (RPA) pilot must monitor a noisy video feed of a crowded marketplace while

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 279–288, 2013.

simultaneously attending to frequent audio and chat communications for task-relevant information, the system could alter the interface to (1) eliminate all the irrelevant information that may clutter the display to reduce the workload demand, and (2) bring into focus central information that needs attention [1]. While different operator states often entail different ideal interface configurations, traditional approaches to UI cannot accommodate this demand.

The feasibility and overall effectiveness of adaptive performance augmentation is dependent on timely and reliable assessments of a human operator's state. The ability to accurately and autonomously define an operator's state, particularly in real-time, has been a much desired yet difficult to achieve capability that has hindered the ability to employ adaptive aiding technologies. One approach that has generated much interest in recent years focuses on the use of physiological data to classify an operator's state. Previous research has shown that physiological measures can be used to detect operator state [2, 3]. Recent improvements in reliability, level of invasiveness, set-up time, and cost of physiological measurement makes it even more compelling. Physiological data also serves as an objective source of information and is theoretically available from any person working in any domain, in contrast to behavioral and situational data which are likely to vary greatly across different work environments.

However, from the perspective of developing an operationally deployable capability for estimating operator states, there have been limitations with regard to: (a) the ability to produce a model with high levels of accuracy across individuals, particularly when the operator state model has not been "trained" to a specific individual; (b) the ability to derive an accurate classification from available real-time data, as opposed to post-hoc analysis in which a much larger spectrum of data are available (e.g., future events, subjective responses, etc.); and (c) the ability to pinpoint the operator's state with high resolution and update frequency. Some of the most successful operator state classification efforts to date have made progress in this endeavor by collecting large sums of data from a specific individual, and subsequently training a custom operator state model for that same individual with machine learning based methods [4]. While this work produced invaluable insights on the possibilities of operator state classification, there are practical limitations to shaping specially trained models to each individual operator using a particular system. More recent work has started to explore cross-subject workload classification [5], however this body of research remains in the early stages. In addition, a prominent theme in the literature to date is the classification of operator states according to very discrete categories, such as "low workload" and "high workload", as well as outputting these categorical state estimates at infrequent intervals. When attempting to employ automated augmentation strategies, the lack of granularity allotted by a "low vs. high" classification and at infrequent update rates may prevent a system from tracking the necessary detailed trends and subtle fluctuations over time that can greatly affect the operator's need for intervention.

In addition, the ideal adaptive augmentation system would be able to incorporate predictions of operator state and its expected impact on human performance. Predictive capabilities would provide an invaluable tool for proactively address problems before they occur. Unfortunately, operator state predictions have not been thoroughly explored, as much of the published research has been focused on historical and

real-time diagnosis of operator state. These predictive capabilities are also held back by the lack of a reliable, continuous, and frequently updated estimate of operator state that supplies the required level of granularity and volume of data necessary to make quality predictions. Collectively, these gaps illustrate the need for a forward-looking approach that can establish an extensible foundation for adaptive aiding strategies; one that is both practical for application and improves the likelihood that dynamic interventions will have a beneficial effect on operator state and job performance.

2 Approach

The objective of our research is to expand upon this existing foundation of research to identify the most relevant and sensitive multi-modal measures of operator states (i.e., neural, physiological, behavioral) and develop algorithms that can assess these states in real time for the purpose of enabling various performance augmentation strategies. In response to this technical challenge, we have designed and implemented an approach that intends to lay a foundation for adaptive aiding technologies to be transitioned to operational system usage.

Our approach relies on innovative physiological-based operator state modeling and classification techniques being formulated and tested within the Air Force Research Laboratory's (AFRL) "Sense, Assess, Augment" taxonomy [6]. To fulfill the "Sense" component of this framework, we have developed a flexible architecture (Figure 1) for collecting and processing physiological, behavioral, and situational data from disparate sources in real-time into a centralized location. The "Assess" component of this framework employs a machine learning based modeling approach that is trained from data sets spanning four categories: Physiological, Self-reported factors, Performance, and Situational. As our test case, the current focus of the assessment component is on operator workload classification as a function of these four categories, given that workload has a demonstrated relationship to task performance and thus is an "augmentable" construct. Lastly, the "Augment" component seeks to "close the loop" on sustained human performance by leveraging the accessibility of real-time continuous workload estimates as the basis for when and how to aid performance. For the purpose of this paper, we focus primarily on the "Sense and Assess" portions of this framework as a stepping stone to achieving the end goal of effective real-time adaptive augmentation strategies.

Our modeling approach is unique on several fronts. First, the inclusion of expansive contextual information to support the model's ability to interpret noisy physiological data has not been substantially explored by other published approaches. We theorize that data characterizing an individual's antecedent health and lifestyle factors, real-time task performance, and situational data from the task environment provide beneficial insight into why physiological patterns occur, thus supporting the ability to "sift through the noise" and ultimately obtain the most meaningful data for operator state classification.

Second, this approach supplies a real-time output with a continuous high-resolution (0-100) scale. We accomplish this by applying machine learning methods to train a

model that identifies the best fit between these available real-time data sources and subjective operator state measurements collected from our experimental paradigm (described in the next section). With respect to our model training approach, we inject noise into each subjective measurement for each corresponding trial to generate an operator state estimate along a continuous scale for model training, under the assumption that few, if any, meaningful operator states are perfectly static over time. Because it is impractical, if not impossible, to obtain operator responses at very frequent intervals (e.g., once per five seconds), it is important to rely on a theoretically-grounded relationship between an available, measurable factor (or set of factors) and the modeled construct of interest as the basis for incorporating noise. The complexity of this component of our approach can range from simple to highly complex depending on the modeled construct and tolerance to error. As an example, for our test case of modeling operator workload, we add noise to self-reported workload ratings based on specially designed algorithms that process contextual data about the situation at each point in time to produce the direction and magnitude of noise.

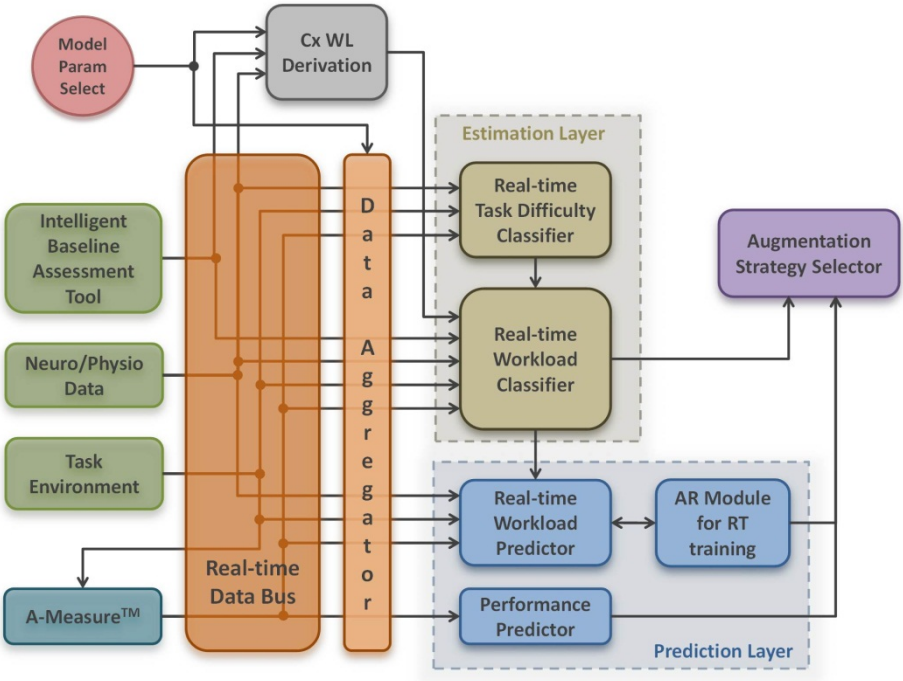

Fig. 1. Data aggregation and modeling architecture for operator state assessment as a foundation for automated adaptive aiding strategies

Third, this approach employs on-line model training capable of improving the precision of the operator state estimation across different individuals over time. The on-line training process is triggered by a scalable set of operator-driven inputs dependent on the state being modeled. Ideally, the scale of these inputs is set to consume the

lowest possible time, effort, attention, and frequency of input from the operator (e.g., 5-second response per every 30 minutes). Using this trigger, a set of sub-components within our architecture dynamically updates the model weights using the operator input data in conjunction with recent physiological, behavioral, and situational data that has occurred during a corresponding timeframe, resulting in more accurate and individualized estimate of operator state that improves over time without the need for a priori custom-built classifiers for each human operator.

Lastly, the predictive layer of this approach utilizes memory of historical data to help facilitate informed, and proactive, augmentation decisions based on expected operator state and performance. The predictive accuracy is, as one would expect, dependent on the level of granularity and update frequency of the real-time operator state classifier. For example, workload estimates on a 0-100 scale and updated once per every five seconds allows a trained model to monitor subtle trends and changes not otherwise possible with highly discrete classifiers (e.g., high versus low); this may potentially be the difference between knowing when, and when not, to intervene with an augmentation strategy. In addition, forecasted knowledge of the situation – such as when a highly tactical and attention-demanding phase of a mission is known to occur – is valuable, if not essential, context that adds to the accuracy of workload and performance predictions.

3 Current Study

3.1 Overview

To develop a prototype operator state model based on this approach, we conducted a model training study at AFRL's Human Universal Measurement and Assessment Network (HUMAN) Laboratory. Our primary objective was to generate data sets that would allow an operator state model of workload to be trained within our defined technical approach. For the scope of this paper, our reporting focuses primarily on model classification accuracy in relation to related published work. Secondary objectives of this study were to validate that subjective workload ratings to be used for training the workload model indeed correspond to the intended task difficulty, and conduct exploratory analysis on the degree to which workload fluctuations correspond to performance fluctuations. These latter objectives are important as a preface to our future research on developing effective augmentation strategies.

3.2 Task Environment

The task environment for this study was based upon a modified version of the Air Force Multi-Attribute Task Battery (AF_MATB) [7]. This PC-based aviation simulation requires an operator to perform an unstable tracking task while simultaneously monitoring warning lights and dials, responding to simulation-generated auditory requests to adjust radio frequencies, and managing simulated fuel flow rates using various key presses. Our rationale for using this task environment was threefold. First, MATB has been used as a testbed to train and develop other models of operator

workload [5], which provides our approach with a benchmark for comparison. Second, MATB allows for linear titration of workload on a high-resolution scale, which provides the necessary task conditions to model beyond "low versus high workload" prior to injecting noise. Third, MATB has long and rich history of research findings that provide a deep understanding of how each task module affects operator workload, as well as the interactions between these factors.

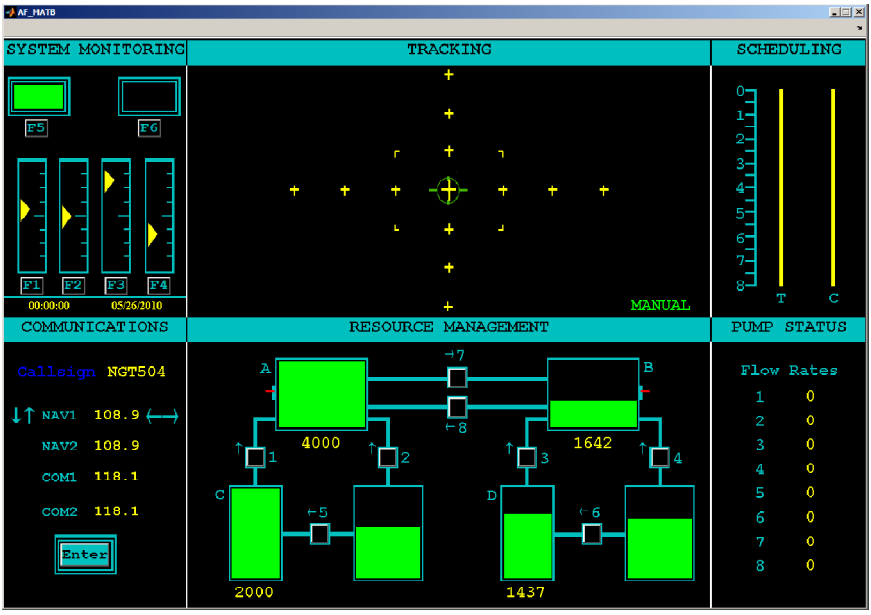

Fig. 2. The operator interface for the AF_MATB task

3.3 Participants

Ten participants served as operators of the AF_MATB system for this study. The only requirement for participation was a familiarity with computer-based systems. Seven participants were male and three participants were female. The age of participants ranged from 23 to 47 years old, with a mean age of 32 years old.

3.4 Experimental Design

Task difficulty was the only independent variable (IV) for this study. We selected task difficulty because this manipulation has been an effective method for inducing varying levels of operator workload [8]. In our attempt to obtain the highest possible level of model granularity, we relied upon 15 levels of task difficulty that intend to linearly span the full range of workload (i.e., low to high). Accordingly, this study employed a one-way experimental design in which a single IV (task difficulty) was manipulated across 15 conditions in order to assess its effects on each dependent variable (DV).

All participants experienced the same 15 conditions; however, the sequence of conditions was counterbalanced to mitigate order effects.

3.5 Dependent Variables

The dependent variables (DVs) were physiological, self-reported, and performance measures collected from participants. Physiological measures included EEG, ECG, and eye-tracking activity (e.g., pupil diameter, fixations, blinks, etc.). Self-reported measures included antecedent lifestyle factors (e.g., demographics, level of exercise, video game experience), recent behavioral factors that can affect physiological state (e.g., sleep quality, current sleepiness, caffeine and food intake), and subjective workload assessments of each condition as measured via the NASA Task Load Index (TLX) scale [9]. Performance measures included primary task performance on the AF_MATB tracking task (distance from centerpoint) and secondary task performance on the lights/gauges task (response time and accuracy).

3.6 Procedures

Each participant went through two sessions: training and data collection. During the training session, participants acquired hands-on training by operating the system during practice scenarios ranging across easy, medium, and hard difficulty conditions. Our goal was to eliminate learning effects during the data collection phase to the extent possible. For the data collection session, participants operated the AF_MATB environment through 15 five-minute scenarios while being monitored with physiological sensors and behavioral data capture software. Each of the 15 scenarios varied by task difficulty and was presented in a quasi-randomized order with five blocks of three scenarios per block. The three blocks in each scenario consisted of a low, medium, and high difficulty block. Physiological sensors collected data on eye movements, blinks, pupil diameter, EEG, and ECG. At the end of each trial, participants completed the NASA TLX questionnaire provided electronically on the AF_MATB.

4 Results and Discussion

4.1 Model Training Results

Within the scope of this study, there are several ways to evaluate the utility of trained model results. First, we evaluated the absolute error (expressed as mean absolute difference percent) between the model's output and the reference continuous workload estimator values upon which the model was trained, which came to an average of 35% for all participants across all trials. For some participants, the average error across trials reached as low as 15%, although other participants produced greater than 50% error. We concluded that while we may have collected many valuable inputs that account for the majority of workload variance for specific individuals, there could be individual differences that were not sufficiently measured.

Second, we analyzed classification accuracy of the trained model when applied retroactively to participant data without providing the model with any direct workload-related input. While categorization is not the ultimate goal of this approach, it is useful as a means for comparing this work to known benchmarks in the literature. Using classification accuracy for low versus high workload, the prototype model produced mean 82.7% accuracy when averaged for entire trials, and 75.7% accuracy on a per five-second basis. We also went a step further by randomly removing two participants from the training set and applying the adjusted model to these removed participants. When averaged for entire trials, the adjusted model produced a mean 87.5% accuracy for low versus high classification for these two participants, and 77.8% on a per five-second basis. When considering our use of continuous high-resolution output as the basis of these classifications – as well as the small sample size and our inclusion of outliers – these results appear to compare favorably to similar work [4, 5]. In addition, our preliminary analysis on the benefits of on-line model training techniques (which are not reported here due to intended scope) has revealed promising trends with regard to additional accuracy generated due to dynamic model weight adjustments over time based on the individual performer.

While the per five-second classification accuracy of our workload model is difficult to empirically validate at this time (i.e., it is not feasible to obtain self-report data every five seconds for comparison), these results provide a quality baseline standard from which to expand our forthcoming work. Our future research will: (a) quantify the benefit of a larger sample size and on-line model training; and (b) identify methods to validate high-resolution output of our approach beyond categorical levels.

4.2 Secondary Analyses

Secondary objectives of this study were to validate that subjective workload ratings to be used for training the workload model indeed correspond to the intended task difficulty, and conduct exploratory analysis on the degree to which workload fluctuations correspond to performance fluctuations. While these findings are not directly related to the formulation of our operator state modeling approach, they can be used as a preface to our future research on developing effective augmentation strategies.

Correlation between participants' self-reported NASA TLX ratings and intended experimental difficulty (1-15) was approximately 0.67, demonstrating that workload was indeed reasonably well connected to the intended task difficulties of scenarios. We further validated this assumption by grouping continuous workload measures used for model training based on intended task difficulty/workload: Low (difficulties 1-5), Medium (difficulties 6-10), and High (difficulties 11-15). Based on these groupings, there was a statistically significant difference between mean continuous workload measures used for model training across each of three groups ($p<0.0001$). Furthermore, we analyzed the addition of our noise injection algorithm to the NASA TLX responses to generate the continuous workload estimates for model training. When averaging the resulting continuous workload estimates across trials, we obtained a correlation of $r = 0.99$ with the actual reported NASA TLX values, which demonstrated the noise injection algorithm did not overly skew workload responses.

Lastly, at an exploratory level we investigated the degree to which the model's estimates of workload provided identifiable clues to when performance decrements might occur. This was an informal analysis done to obtain a realistic expectation as to how frequently performance decrements could be identified proactively, using workload as the "leading indicator" and/or "trailing indicator" of their occurrence. The example illustration in Figure 3 demonstrates one recurring trend in which a performance decrement can serve as a leading indicator of workload spikes, followed by subsequent behavioral changes in reaction to these effects. Currently, we are quantitatively formalizing the complex relationships between workload and performance as a precursor to intelligent augmentation strategy selection in real-time mission settings.

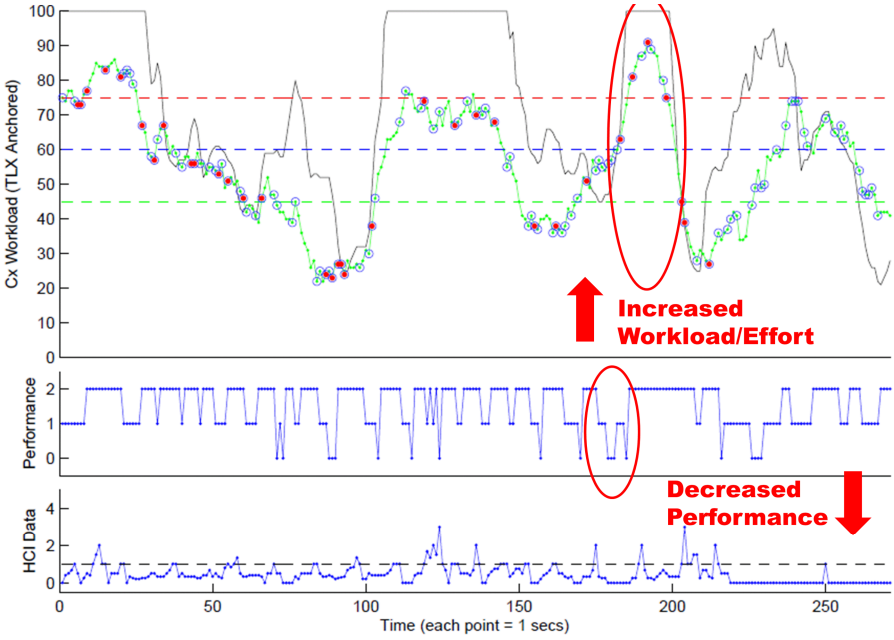

Fig. 3. Example of a workload spike as a leading indicator for a performance decrement

5 Conclusions

This paper described a novel technical approach for establishing real-time estimates of operator states on a continuous, high-resolution scale for the purpose of improving the ability to employ effective adaptive aiding strategies for performance augmentation. Using operator workload as a test case, our research to date has served as a key stepping stone with regard to establishing a level of accuracy in line with the published state of the art. Future research will focus on improving model accuracy through additional data collection, optimizing components of the model architecture (e.g., on-line training), and additional measures that may account for a larger percentage of workload variance. A critical next step is also the design of a model validation

paradigm that enables empirical investigation of workload estimation accuracy on a continuous 0-100 scale. Finally, we will quantitatively represent the complex relationships between workload and performance, which may provide substantial benefit to the employment of automated aiding strategies to mitigate performance decrements.

Acknowledgement. This material is based upon work supported by the Air Force Research Laboratory (AFRL) under Contract No. FA8650-11-C-6236. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of AFRL. The authors would like to thank: Seamus Sullivan, Noah DePriest and Zachary Zuzack for software support; Matt Middendorf, Michael Hoepf, Cassandra Christman, and Chelsey Credlebaugh for data collection support; and Justin Estepp for feedback on our technical approach.

Distribution A: Approved for public release; distribution unlimited. 88ABW Cleared 4/02/2013; 88ABW-2013-1591.

References

1. Parasuraman, R.: Neuroergonomics: Brain, cognition, and performance at work. Current Directions in Psychological Science 20, 181–186 (2011)
2. Wilson, G.F., Eggemeier, F.T.: Physiological measures of workload in multi-task environments. In: Damos, D. (ed.) Multiple-task Performance, pp. 329–360. Taylor & Francis, London (1991)
3. Schnell, T., Keller, M., Macuda, T.: Application of the Cognitive Avionics Tool Set (CATS) in Airborne Operator State Classification. In: Augmented Cognition International Conference, Baltimore, MD (2007)
4. Wilson, G.F., Russell, C.A.: Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks. Human Factors 45(4), 635–644 (2003)
5. Wang, Z., Hope, R.M., Wang, Z., Ji, Q., Gray, W.: Cross-subject workload classification with a hierarchical Bayes model. NeuroImage 59(1), 64–69 (2012)
6. Galster, S.: Sense-Assess-Augment: A Taxonomy for Human Effectiveness. In: Proceedings of the Seventeenth International Symposium on Aviation Psychology. Dayton, OH (in press)
7. Miller, W.D.: The U.S. Air Force-developed adaptation of the Multi-Attribute Task Battery for the assessment of human operator workload and strategic behavior (Tech. Rep. No. AFRL-RH-WP-TR-2010-0133) (2010)
8. Backs, R.W., Seljos, K.A.: Metabolic and cardiorespiratory measures of mental effort: The effects of level of difficulty in a working memory task. International Journal of Psychophysiology 16, 57–68 (1994)
9. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In: Peter, A.H., Najmedin, M. (eds.) Advances in Psychology, vol. 52, pp. 139–183. North-Holland, Amsterdam (1988)