Workload-Based Assessment of a User Interface Design

Patrice D. Tremoulet[1], Patrick L. Craven[1], Susan Harkness Regli[1], Saki Wilcox[1], Joyce Barton[1], Kathleeen Stibler[1], Adam Gifford[1], and Marianne Clark[2]

[1] Lockheed Martin Advanced Technology Laboratories
3 Executive Campus, Suite 600, Cherry Hill, NJ, USA
{polly.d.tremoulet,patrick.craven,susan.regli,
sakirenecia.h.wilcox,joyce.h.barton,kathleen.m.stibler,
adam.gifford}@lmco.com
[2] 2001 South Mopac Expressway, Suite 824, Austin, TX, USA
poythressm@aol.com

Abstract. Lockheed Martin Advanced Technology Laboratories (LM ATL) has designed and developed a tool called Sensor-based Mental Assessment in Real Time (SMART), which uses physiological data to help evaluate human-computer interfaces (HCI). SMART non-intrusively collects and displays objective measures of cognitive workload, visual engagement, distraction and drowsiness while participants interact with HCIs or HCI prototypes. This paper describes a concept validation experiment (CVE) conducted to 1) demonstrate the feasibility of using SMART during user interface evaluations and 2) validate the EEG-based cognitive workload values derived from the SMART system by comparing them to three other measures of cognitive workload (NASA TLX, expert ratings, and expected workload values generated with Design Interactive's Multimodal Information Decision Support tool). Results from the CVE indicate that SMART represents a valuable tool that provides human factors engineers with a non-invasive, non-interrupting, objective method of evaluating cognitive workload.

Keywords: Cognitive workload, human computer interaction, human factors, usability, evaluation, user interface design.

1 Introduction

In 2005 and 2006, the Office of Naval Research (ONR) Disruptive Technologies Opportunity Fund supported Lockheed Martin Advanced Technology Laboratories' (LM ATL) research effort exploring the use of neuro-physiological data to measure cognitive workload during a human-computer interface (HCI) evaluation. As a part of this effort, LM ATL designed and developed a tool called Sensor-based Mental Assessment in Real Time (SMART), which non-intrusively collects physiological data (electroencephalographs (EEG), heart rate variability (HRV), galvanic skin response and pupil size) from subjects while they interact with HCIs or HCI prototypes. SMART uses the data to derive objective measures of cognitive workload, visual engagement, distraction and drowsiness, which may be used to evaluate the efficacy of design alternatives, e.g., by helping to identify events of interest during

V.G. Duffy (Ed.): Digital Human Modeling, HCII 2009, LNCS 5620, pp. 333 – 342, 2009.

usability tests, thus reducing data load and providing timely evaluation results to suggest design changes or to validate design.

SMART's sensor-based measure of cognitive workload offers several advantages over existing workload measures, including: 1) increased precision in measuring the subject's cognitive state via a moment-by-moment data collection, 2) obtaining objective measurement of cognitive workload while the test is being performed, and 3) not distracting subjects from their primary task (e.g., by interrupting to collect subjective ratings or requiring them to attend to a secondary task). However, SMART's cognitive workload measure needed to be validated, so LM ATL conducted a concept validation experiment (CVE) to demonstrate the feasibility of using SMART during user interface evaluations as well as to collect data necessary to validate the sensor-based cognitive workload values derived from the SMART system by comparing them to NASA TLX, expert ratings, and Multimodal Information Decision Support (MIDS) expected values.

In most respects, the CVE was similar to a traditional usability study. Twelve sailors interacted with a high-fidelity prototype of a future release of the Tactical Tomahawk Weapons Control System (TTWCS), performing tasks required to execute missile strike scenarios. However, there were two major differences. First, the CVE scenarios were designed not only to be operationally valid but also to include discrete phases that require specific levels of workload. Moreover, while interacting with the prototype, participants in the CVE wore a set of neurological and physiological sensors including a wireless continuous EEG and Electrocardiogram (EKG) sensor, wired EKG (the wired EKG was used as a backup for the newer wireless configuration) and galvanic skin response (GSR) sensors, and an off-head, binocular eye tracker that logs point of gaze and pupil diameters.

1.1 Sensor-Based Mental Assessment in Real-Time (SMART)

SMART provides critical support for interpreting physiological data collected during usability evaluations. SMART logs and displays system events, physiological responses, and user actions while study participants continue to interact with system of interest, in this case a prototype of a future TTWCS system. Advances in neuro-technology and physiological measurements have enabled information to be captured that helps identify and indicate psychological states of interest (e.g., boredom, overloaded, and engaged), which aid and human factors engineers in the evaluation of an HCI.

SMART provides four critical types of information during usability evaluations:

- *Real-Time Logging:* Experimenters have the ability to enter events into the log to represent significant events that are not automatically logged. Prior to testing, the experimenter can set up events of interest with a quick key identifier so that expected events can be manually logged during testing.
- *Real-Time Monitoring:* During testing, experimenters can log events and monitor physiological sensors data (Fig. 1).
- *Time Synchronization:* Time synchronization between the physiological sensors logs and the test system logs is crucial in accurately matching sensor derived events

to test platform and participant driven events. After testing, the experimenter can view data via the Timeline summary (Fig. 2).

- *Data Extraction:* Data is also extracted and presented in a CSV format, suitable for uploading into standard statistical analysis applications.

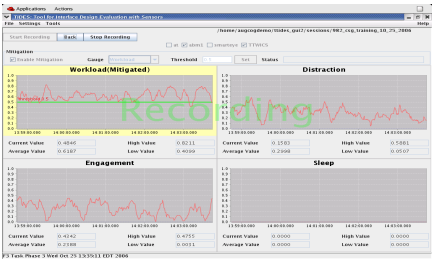

Fig. 1. Real-time monitoring

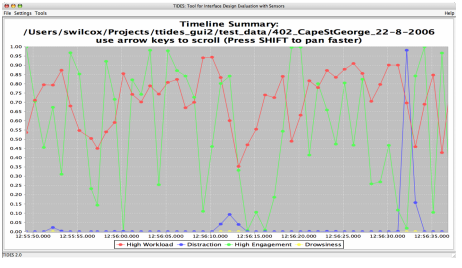

Fig. 2. Timeline Summary

1.2 SMART's Cognitive Workload Measure

Lockheed Martin Advanced Technology Laboratories (LM ATL) worked with Advanced Brain Monitoring (ABM) to develop an Electroencephalogram (EEG)-based gauge of cognitive workload [2] [4]. LM ATL employees wore ABM's EEG acquisition system while performing a variety of classic experimental psychology tasks in which working memory load was varied (e.g., by varying the number of items that needed to be remembered in an N-back task). The EEG acquisition system measured the electrical activity of the participants' brains with sensors placed on the scalp, allowing data to be captured unobtrusively. The EEG signals reflect the summated potentials of neurons in the brain that are firing at a rate of milliseconds.

Discriminant function analyses determined appropriate coefficients for linearly combining measures derived from continuous EEG into a cognitive workload index, which ranges from 0 to 1.0 (representing the probability of being classified as "high workload"). The index values were validated against an objective appraisal of task difficulty and subjective estimates of workload and task focus.

The cognitive workload index derived from this research effort increases with increasing working memory load and during problem solving, mental arithmetic,

integration of information, and analytical reasoning and may reflect a sub-set of executive functions. The cognitive workload levels from this index are significantly correlated with objective performance and subjective workload ratings in tasks with varying levels of difficulty including forward and backward digit span, mental arithmetic and N-back working memory tests [4] [3].

2 Method

A Concept Validation Experiment (CVE), patterned after an HCI usability study in which active-duty personnel run operationally valid scenarios designed such that different phases elicit different levels of cognitive workload, was conducted in Norfolk, VA, from August 22-31, 2006.

2.1 Participants

Average age of the twelve active-duty participants was 29, and their years of service ranged from 1-18 years. Six participants had experience with one of the previous Tomahawk systems. All but one had some experience at Tomahawk workstations (averaging over six years) and had participated in an average of 52 Sea-Launched Attack Missile Exercises (SLAMEX) and 28 theatre exercises. Two participants had experience participating in an operational launch. All participants were male; one was left-handed, and two reported corrected 20/20 vision.

2.2 Equipment and Materials

Testing was conducted at offices at the Naval Station in Norfolk, VA. The TTWCS prototype was simulated on a Dell Inspiron 8500 laptop connected to two, 19-inch flat panel LCD monitors displaying the prototype software. The monitors were positioned vertically to re-create the configuration of the actual Tomahawk workstation. The prototype system recorded timestamps for system and user events to a log file.

One video camera was located behind the participant, and video was captured on digital video tape. The output from the console monitors was sent through splitters to two 19-inch displays to allow observers to view the screen contents from a less intrusive location.

Continuous EEG data was acquired from a wireless sensor headset developed by ABM using five channels with the following bi-polar montage: C3-C4, Cz-POz, F3-Cz, Fz-C3, Fz-POz. Bi-polar differential recordings were selected to reduce the potential for movement artifacts that can be problematic for applications that require ambulatory conditions in operational environments. Additionally, EKG information was collected on a sixth channel. Limiting the sensors (six) and channels (five) ensured the sensor headset could be applied within 10 minutes. The EEG data were transmitted via a radio frequency (RF) transmitter on the headset to an RF receiver on a laptop computer containing EEG collection and processing software developed by ABM. This computer was then linked to an Ethernet router that sent the EEG data to SMART.

Wired galvanic skin response (GSR) and EKG sensors were connected to a ProComp™ Infiniti conversion box, which transmitted the signal via optical cable to

two destinations. The first destination was a laptop containing Cardio Pro™ software, and to the SMART laptop where software logged the GSR data and used the EKG signal to calculate heart rate variability (HRV).

Eye gaze and pupil dilation data were collected using the SmartEye™ system in which two cameras were positioned on either side of the monitor. An infrared emitter near each camera allowed the SmartEye™ software to calculate the associated eye movement and dilation data.

Paper questionnaires included a background questionnaire and a user satisfaction questionnaire. An electronic version of the NASA TLX was administered on the TTWCS task machine in between task phases.

The test administrator recorded notes concerning the testing session and completed rating of the effort observed by the participant during the test session. Subsequent to the test session, aspects of the scenarios and individual user behavior were coded and entered into the MIDS tool. This tool produces a second-by-second total workload value for the entire scenario.

2.3 Experimental Design

The SMART workload validation was a correlational design in which the validity of a novel measure (sensor-based cognitive workload) would be compared with other measures that typify accepted practices for measuring workload within the HCI and Human Factors engineering research communities.

2.4 Procedure

Each participant's test session lasted approximately eight hours. The participants were briefed on the purpose and procedures for the study and then completed a background questionnaire with information regarding demographics, rank, billet, and Tomahawk experience. After the questionnaire was completed, a cap was placed on the participant's head and six EEG sensors were positioned in it (at F3, Fz, C3, Cz, C4, and Pz). Two EKG sensors were placed on the right shoulder and just below the left rib. The participant was instructed to tell the testing staff if they were uncomfortable before or during the experimental session. An impedance check was done to ensure that interference to the signals was at 50 ohms or below; once this was verified, the RF transmitter and receiver were turned on.

The participant was then asked to perform three tasks over 30 minutes to establish an individual EEG baseline and calibrate the sensors. The tasks required reading instructions and performing basic visual and auditory monitoring tasks. Next, individual profiles were created for the SmartEye™ eye tracking system by taking pictures while the participant looked at five predetermined points on the displays. Additionally, during hands-on training, specified facial features were marked on the facial images to further refine the participant's profile.

Finally, three EKG sensors were attached, one on the soft spot below each shoulder and one on the center of the abdomen, and two GSR sensors were placed on the second and fourth toes such that the sensor was on the "pad" of the toe. A towel was wrapped around the participant's foot to keep it from getting cold.

Table 1. Sensor devices used in CVE

Sensor Vendor	Description	Hardware
ABM	Collects EEG data from five channels and EKG on a sixth channel	Six EEG electrodes Two EKG electrodes ABM sensor cap RF transmitter and receiver
SmartEye™	Collects point-of-gaze and pupilometry data	Two cameras Two infrared-flashes
Thought Technology	Collects EKG and GSR data HRV calculated through third-party software	ProComp Infiniti converter Two GSR electrodes Three EKG electrodes

Once all sensors were applied and were collecting data, SMART software was started and began collecting data. Software from ABM, SmartEye™, and CardioPro™ were used throughout the experiment to monitor the signals from the sensors (Table 1). Network time protocol (NTP) was used to ensure that all machines were synchronized to within a fraction of a millisecond.

Once the sensor equipment was calibrated and system software was running, participants were trained through lecture and given hands-on practice using the prototype. The lecture portion of training consisted of PowerPoint slides, which gave an overview of the system, as well as detailed pictures and descriptions of the windows and interactions that would be part of the participant's experimental task.

During the hands-on practice, a member of the testing staff guided the participant through sample tasks. The participant performed one practice trial using the system, during which various scenario tasks were accomplished. The testing staff answered any questions that the participant had during training. Training took one and a half hours (one hour of lecture and thirty minutes of hands-on practice). After training, participants were given a thirty-minute break.

During the experimental test session, the participants were presented with a scenario that included various tasks associated with preparing and launching Tomahawk missiles, similar to those presented during training. The participant was asked to respond to system events and prompts. An experimenter observed and recorded errors and additional observations using SMART and supplemental written notes. SMART logged the sensor data and the objective cognitive measures derived from it, and the TTWCS prototype logged user and system events throughout the test session. The sessions were also video and audio recorded. The video showed only the back of the participant's head and the two computer screens.

In the test scenario, participants were asked to execute two, 10-missile salvos from receipt of the strike package until the missiles launched for first strike package. The main criterion for successful human-machine performance in this cognitive task environment was the degree to which missiles were successfully launched at their designated launch time (T=0).

The scenario took approximately one hour and 15 minutes to complete. The scenario was divided into four task phases (each desired to have a different level of workload) that occurred in a specific sequence for a strike:

- *Phase 1:* Initial validation with no error and engagement planning with errors
- *Phase 2:* Validation due to receipt of execute strike package, preparation for execution, and emergent targets
- *Phase 3:* Monitor Missiles
- *Phase 4:* Launch missiles with emergent targets and receive and prepare second strike package

At the end of each phase, the prototype was paused and the participant asked to fill out a questionnaire on perceived workload (NASA TLX). Then the scenario was resumed.

After the fourth phase, the sensors were removed and the participant was asked to fill out a questionnaire on the perceived satisfaction with the system and then was debriefed to discuss any questions that they had.

2.5 Data Preparation

Four measures of cognitive workload were collected during the test sessions:

- *Sensor cognitive workload (EEG-based cognitive workload value):* Scores from the neurophysiologically-based gauges (CW-EEG) were generated by taking logs of second-by-second gauge output and averaging the workload values during the four different task phases. Data points associated with noisy sensor readings were removed before analysis ([1] provides a description of this procedure.)
- *NASA TLX (generated by participant survey):* Total workload scores from the NASA TLX were used in the analyses. The total workload score is comprised of weighted scores of the six subscales (mental demand, physical demand, temporal demand, performance, effort, and frustration). One participant's data was removed because he rated all four task phases in both scenarios the same.
- *MIDS (generated by expert observation and cognitive conflict algorithm):* Scores for the MIDS measure were generated for approximately half the participants. Close examination of the task domain and videotape of the study allowed for generation of estimates of workload for individual sensory channels (i.e., visual, auditory, and haptic), cognitive channels (i.e., verbal, spatial) and response channels (i.e., motor, speech). Workload was calculated based on a task timeline by summing (1) the purely additive workload level across attentional channels, (2) a penalty due to demand conflicts within channels, and (3) a penalty due to demand conflicts between channels. The amount of attention the operator must pay to each channel in the performance of each task used a 5-point subjective rating scale (1: very low attentional demand; 5: very high attentional demand). These estimates were combined to create an overall cognitive workload estimate for each second that the participant was completing the scenarios ([5] provides details of this technique.) Values from the six participants were averaged and correlations were made with average values of the other measure.
- *Expert Rating (generated by expert observation):* The expert ratings were generated by three HCI experts using a seven-point Likert scale. Experts rated the participants on six dimensions of interaction with the task. The mental effort rating was then extracted and scaled by rater to help eliminate individual rater tendencies.

3 Results

As described above, four measurements were collected during each of the four phases of the test scenario. The measures estimated workload using either an external observer who monitored the actions of the participant, by the participant himself, or through the use of physiological-sensor-based gauges. The results are presented below first in terms of descriptive statistics and then in terms of correlations of the measures.

3.1 Descriptive Statistics

Table 2 shows the mean values for the four measures. For all four measures, Phase 4 had the highest mean value. During this phase, participants were required to launch missiles while simultaneously preparing a second strike package. They also were given emerging targets, requiring quick response. For CW-EEG, Expert Rating and MIDS, Phase 3 was the lowest mean value. During this phase, the participants were primarily performing a vigilance task as the missiles were being prepared for launch, which required no specific interaction with the TTWCS HCI prototype.

Table 2. Overall descriptive statistics

Metrics	Phase	Mean
CW-EEG (0 to 1.0)	Phase 1	0.703
	Phase 2	0.702
	Phase 3	0.659
	Phase 4	0.708
NASA TLX (0 to 100)	Phase 1	22.56
	Phase 2	28.62
	Phase 3	24.64
	Phase 4	41.56

Metrics	Phase	Mean
Expert Rating (1 to 7)	Phase 1	2.75
	Phase 2	3.46
	Phase 3	2.04
	Phase 4	4.78
MIDS (1 to 5)	Phase 1	11.31
	Phase 2	11.11
	Phase 3	3.54
	Phase 4	13.31

3.2 Validation of EEG-Based Cognitive Workload

Validation of the EEG-based cognitive workload index (CW-EEG) was performed by computing correlations among the scores (NASA TLX, expert rating) or average scores (CW-EEG, MIDS) across subjects of the four measures collected during the CVE. Table 3 lists the Pearson's product-moment correlation coefficient (r) of four measures of cognitive workload. Note that MIDS data was only coded on half the participants' data).

Table 3. Workload measures correlations table

Metrics	CW-EEG	NASA-TLLS	Expert	MIDS
CW-EEG	--	0.19^	0.38**	0.51**
NASATLX	0.19^	--	0.34**	0.40*
Expert Ratings	0.38**	0.34**	--	0.56**
MIDS	0.51**	0.40*	0.56**	--

***p*<.001, **p*<.01, ^*p*<.10.

As the table indicates, CW-EEG is significantly related to both the expert rating and the MIDS cognitive workload estimate. The probability that CW-EEG was related to the NASA TLX estimate of workload was not significant in the examined population.

4 Discussion

The results of this validation study are encouraging for continued advancement in our understanding of users' cognitive workload. The desire to simultaneously increase the military's capabilities while reducing staffing requirements has resulted in greater demands on today's military system operators. The capability to identify the cognitive demands of tasks has the potential to identify the limitations of existing technology interfaces and highlight those periods of interaction that result in excessive cognitive demand for the operator. Traditional methods for assessing cognitive workload have serious limitations such as intrusiveness and bias. For example, the NASA TLX requires administering a survey that demands time and attention from study participants, and the results of this survey fail to discriminate between long and short time periods of testing. Moreover, the NASA TLX includes physical demand as one of its dimensions and physical demands for C2 tasks have low variation. On the other hand, if an HCI expert evaluates the operator during a task, there is the potential for bias. Experimenters can only observe the expression of operators and must estimate the workload based on aspects of the task coupled with the particular reactions of the participant. Administering secondary tasks can produce an objective assessment, but the measures produced by this method are extremely sensitive to variations in either primary or secondary tasks. In contrast, SMART's EEG-based measure provides an unbiased estimate that has minimal intrusiveness during the testing period. Additionally, high-resolution scores can be obtained for the duration of the test period, and this data can later be aggregated using a variety of methods.

The study presented here was designed to assess the valicity of SMAR's workload measure, by comparing it to three alternative methods for measuring cognitive workload. The results of this study indicate that the novel physiologically-based measure is significantly correlated with the expert rating and the MIDS workload estimate. The relation between SMART's EEG-based workload index and the NASA TLX was not significant in the examined population. One possible explanation for the lack of a significant relation between SMART's workload index and the NASA TLX is that the third-party perspective of the expert rating and MIDS may have been better related to the sensor-based gauge since all of these represent more objective measurements that do not include the participants' individual preferences and desires. A second possible explanation is the validated measure of workload for the NASA TLX includes six weighted subscales; one would only expect SMART's workload measure to be related to one of these subscales (mental demand). Depending on the participants' particular weighting of the subscales, the relation between SMARTS workload index and the NASA TLX may have been obscured. Additional studies are needed to more closely examine the relation between user ratings, such as the NASA TLX, and physiologically-based measures of cognitive workload.

5 Conclusion

The concept validation experiment discussed in this paper provides empirical validation of the utility of physiological monitoring as a method for non-invasively evaluating the workload associated with performing tasks using particular user interface designs. This experiment successfully demonstrated that LM ATL's SMART tool can produce cognitive workload, visual engagement, distraction and drowsiness measures with high frequency without any action required by the usability study participant. Moreover, analyses of the experiment data revealed significant correlations between the average of the cognitive workload index values produced by SMART, expert ratings, and MIDS expected workload values. These results, taken together, indicate that SMART represents a valuable tool for user interface designers by providing a non-invasive, non-interrupting, objective, real-time, domain-independent, task-independent cognitive workload measures.

Acknowledgments. This research was supported by Office of Naval Research program "Disruptive Technology Opportunity Fund" via the Space and Naval Warfare Systems Command (SPAWAR). We thank the Program Management Activity (PMA) 280 Project Office and the TTWCS System Development Activity (SDA) for their support of this project. We also thank Gene Kocmich of Northrop Grumman for his support in arranging the testing facilities in Norfolk, and Bill Fitzpatrick and Kevin Cropper of Johns Hopkins University's Applied Physics Laboratory for their assistance running the concept validation experiment described here.

References

1. Berka, C., Levendowski, D., Cventinovic, M., Petrovic, M., Davis, G., Lumicao, M., Zivkovic, V., Popovic, M., Olmstead, R.: Real-time Analysis of EEG Indices of Alertness, Cognition, and Memory with a Wireless EEG Headset. International Journal of Human-Computer Interaction 17(2), 11–170 (2004)
2. Berka, C., Levendowski, D.J., Ramsey, C.K., Davis, G., Lumicao, M.N., Stanney, K., Reeves, L., Regli, S., Tremoulet, P.D., Stibler, K.: Evaluation of an EEG-Workload Model in an Aegis Simulation. In: Caldwell, J.A., Wesensten, N.J. (eds.) Biomonitoring for Physiological and Cognitive Performance during Military Operations. Proceedings of the International Society for Optical Engineering, vol. 5797, pp. 90–99 (2005)
3. Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Zivkovic, V.T., Olmstead, R.E., Tremoulet, P.D., Craven, P.L.: EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning and Memory Tasks. Aviation, Space, and Environmental Medicine 78(5), Section II (2007)
4. Craven, P., Belov, N., Tremoulet, P., Thomas, M., Berka, C., Levendowski, D., Davis, G.: Cognitive Workload Gauge Development: Comparison of Real-time Classification Methods. In: Schmorrow, D., Stanney, K., Reeves, L. (eds.) Foundation in Augmented Cognition, 2nd edn., pp. 66–74. Strategic Analysis, Inc., Arlington (2006)
5. Hale, K., Reeves, L., Samman, S., Axelsson, P., Stanney, K.: Validation of predictive workload component of the multimodal information design support (MIDS) system. In: Proceedings of the 49th Annual Human Factors and Ergonomic Society Meeting, Orlando, FL, September 26-30 (2005)