Quantification of Trainee Affective and Cognitive State in Real-time

Christina Kokini, Meredith Carroll, Ruben Ramirez-Padron, Kelly Hale
Design Interactive, Inc.
Oviedo, FL
Christina, Meredith, Ruben, Xuezhong, Kelly@designinteractive.net

Robert Sottilare, Benjamin Goldberg
U.S. Army Research Laboratory
Human Research and Engineering Directorate
Orlando, FL
Robert.Sottilare, benjamin.s.goldberg @us.army.mil

ABSTRACT

Intelligent Tutoring Systems (ITS) have yet to reach training effectiveness levels rivaling those of human tutors, partially due to their inability to recognize and adapt to trainee cognitive and affective states. While many studies have examined expensive sensor suites to capture physiological indicators of cognitive and affective states, the authors' previous work presented an innovative conceptual framework for utilizing low-cost sensors to capture specific states in real-time. Such measures are expected to improve an ITS's ability to automatically adapt to a trainee's readiness to learn.

The current set of two experiments aimed to develop real-time classifiers for six distinct affective and cognitive states (anger, fear, boredom, workload, engagement, distraction) utilizing low-cost, non-invasive (neuro)physiological and behavioral sensors. In the first experiment, participants completed a within-subjects, repeated-measures study in which the independent variable was task type - each task was designed to induce a subset of the targeted states. Dependent variables theorized to indicate targeted states included heart rate, postural sway, pupil diameter, and electroencephalography (EEG) band activity. Each metric was captured via low-cost sensor technology. Validated, ground-truth measures of targeted cognitive and affective states were captured via a 10-channel EEG headset and associated algorithms, and a subjective emotional rating tool, respectively. Several challenges were encountered with the low-cost sensors, including limitations in sensitivity to physiological changes and reliability of data collection. Small design and procedural changes were made for the second experiment, and good logistic regression classifiers for the affective states of boredom and fear were obtained. Additionally, logistic model trees showed good generalization capability when validated as classifiers for the cognitive states. This paper presents study results, lessons learned and implications for future research.

ABOUT THE AUTHORS

Christina Kokini is a Research Associate at Design Interactive, Inc., and has been involved in design, development and evaluation of virtual training tools for the Office of Naval Research, and the Army Research Laboratory. Her work focuses on training system design, development, and usability, including designing functionality into products to support user requirements, as well as conducting usability and training effectiveness evaluations. She holds a Master's degree from Penn State University in Industrial Engineering with a Human Factors Option, where her research focused on the direct effect of contextual characteristics on the perceived usability of a product. She also has a Bachelor's degree from Purdue University in Industrial Engineering.

Meredith B. Carroll, PhD is a Senior Research Associate at Design Interactive, Inc. and has been involved in design, development and evaluation of performance assessment tools and virtual training tools for the Office of Naval Research, the Air Force Research Laboratory, and the Army Research Laboratory. Her work focuses primarily on individual and team performance assessment, including physiological and behavioral measurement, performance diagnosis, and training remediation through feedback and training adaptation. She has also performed extensive work conducting task analyses, designing virtual training environments and performance assessment tools and conducting training effectiveness evaluations. Her research has focused on human/team performance and training in complex systems in aviation and military domains, with focuses on perceptual skills and decision making. She received her B.S. in Aerospace Engineering from the University of Virginia, her M.S. in Aviation

Science from Florida Institute of Technology and her Ph.D. in Applied Experimental and Human Factors Psychology from the University of Central Florida.

Ruben Ramirez-Padron is a Software Engineer II at Design Interactive, Inc. He has a research background in novelty detection, statistical modeling, kernel methods for pattern analysis, and online machine learning. At Design Interactive, Inc., he has been involved in the proposal and validation of intelligent data analysis methods and machine learning techniques for several R&D projects. He received his Bachelor's degree in Computer Science from "Universidad Central de Las Villas", Cuba, in 1996. He graduated with a Master's degree in Computer Engineering (Intelligent Systems Track) from the University of Central Florida (UCF) in 2009. Currently, he is a PhD candidate at the Department of Electrical Engineering and Computer Science at UCF.

Kelly Hale, PhD is Sr. Vice President of Technical Operations at Design Interactive, Inc., and has over 12 years experience in human systems integration research and development. Her R&D in the areas of augmented cognition, multimodal interaction, training sciences, and virtual environments has been funded by the Defense Advanced Research Projects Agency (DARPA), Office of Naval Research (ONR), Intelligence Advanced Research Projects Activity (IARPA), and Department of Homeland Security (DHS), as well as other sources. Through these efforts, Kelly and her team have developed advanced neurophysiological measurement techniques, including a patent-pending Fixation-locked Event-Related Potentials (FLERPs) approach to capture electroencephalography ERP data in a naturalistic setting, and have advanced real-time mitigation strategy framework and induction techniques to optimize training, situation awareness, and operational performance through optimization of user cognitive and physical state. She received her BSc in Kinesiology/Ergonomics Option from the University of Waterloo in Ontario, Canada in 1999, and her Masters and PhD in Industrial Engineering, with a focus on Human Factors Engineering, from the University of Central Florida in 2001 and 2006, respectively.

Robert A. Sottilare, PhD is the Associate Director for Science & Technology within the U.S. Army Research Laboratory - Human Research & Engineering Directorate (ARL-HRED). Dr. Sottilare has over 25 years of experience as both a U.S. Army and Navy training & simulation researcher, engineer and program manager. He leads the Simulation and Training Technology Center's international program, and participates in training technology panels within both The Technical Cooperation Program (TTCP) and NATO. He has a patent for a high resolution, head mounted projection display (U.S. Patent 7,525,735) and his recent publications have appeared in the Journal for Defense Modeling and Simulation, the NATO Human Factors and Medicine Panel's workshop on Human Dimensions in Embedded Virtual Simulation and the Intelligent Tutoring Systems Conference. Dr. Sottilare is a graduate of the Advanced Program Managers Course at the Defense Systems Management College at Ft. Belvoir, Virginia and his doctorate in modeling & simulation is from the University of Central Florida. The focus of his current research program is in machine learning, trainee modeling and the application of artificial intelligence tools and methods to adaptive training environments.

Benjamin Goldberg is a member of the Learning in Intelligent Tutoring Environments (LITE) Lab at the U.S. Army Research Laboratory's (ARL) Simulation and Training Technology Center (STTC) in Orlando, FL. He has been conducting research in the Modeling and Simulation community for the past 4 years with a focus on adaptive learning and how to leverage Artificial Intelligence tools and methods for adaptive computer-based instruction. Currently, he is the LITE Lab's lead scientist on instructional strategy research within adaptive training environments. Mr. Goldberg is a Ph.D. student at the University of Central Florida and holds an M.S. in Modeling & Simulation. Prior to employment with ARL, he held a Graduate Research Assistant position for two years in the Applied Cognition and Training in Immersive Virtual Environments (ACTIVE) Lab at the Institute for Simulation and Training.

Quantification of Trainee Affective and Cognitive State in Real-time

Christina Kokini, Meredith Carroll, Ruben Ramirez-Padron, Kelly Hale
Design Interactive, Inc.
Oviedo, FL
Christina, Meredith, Ruben, Kelly@designinteractive.net

Robert Sottilare, Benjamin Goldberg
U.S. Army Research Laboratory
Human Research and Engineering Directorate
Orlando, FL
Robert.Sottilare, benjamin.s.goldberg @us.army.mil

INTRODUCTION

Current intelligent tutoring systems (ITSs) strive to provide one function regardless of the domain: to maximize learning outcomes by tailoring instructional/training content to strengths and weaknesses associated with a given learner. The traditional approach to ITS implementation is assessing user interactions against performance-based models to determine deficiencies that require guidance or further instruction. Such approaches have been found to be effective in well-defined domains where misconceptions and deviations from desired performance are easily determined (Graesser et al., 2005; Stottler et al., 2001; VanLehn et al., 2005). However, ITSs that adapt solely based on performance ignore the influence that a learner's affective and cognitive states have on learning and retention. For adaptive training systems to reach full potential, technologies need to be in place that track physiological and behavioral markers linked to affective and cognitive states shown to impact learning (see Carroll et al., 2011 for detailed list). This enables a system to monitor reactive tendencies to training stimuli and to determine an individual's readiness to learn (Stevens, Galloway, and Berka, 2007). This information can then be used to inform adaptations to system elements intended for maintaining optimal learning states.

Addressing this functional gap is not a new idea. The past decade has seen a number of studies examining sensor technologies for the purpose of informing state representations (Burleson and Picard, 2004), and remains a current thrust within the ITS and affective computing research communities (Calvo and D'Mello, 2010). However, the inherent problem with work in this field is applying these tools on a large scale. Sensor technologies are often expensive, making the integration of such tools into existing ITS platforms unreasonable. To address this limitation, a conceptual framework for applying low-cost sensors to capture real-time physiological and behavioral markers for assessing state variables has been developed (Carroll et al., 2011). The goal of the current effort was to determine if correlations exist between low-cost sensor metrics and associated ground-truth measures of targeted cognitive and affective states, and to determine if such low-cost sensors could accurately and reliably measure distinct cognitive and affective states. This paper presents a series of studies evaluating the efficacy of utilizing low-cost sensor solutions for detecting trainee cognitive and affective states for use in informing state sensitive ITS student models.

Three affective and three cognitive states were selected for inclusion in the studies based on 1) their impact on learning, and 2) their potential to be measured with low-cost (neuro)physiological and behavioral sensors. The affective states chosen were anger, fear, and boredom, which have all been found to have negative impacts on learning (e.g. McQuiggan et al., 2007; D'Mello et al., 2007), and which have been shown to be correlated with physiological data such as heart rate and posture (Lisetti and Nasoz, 2004; Woolf et al., 2009). The cognitive states chosen were engagement, distraction, and workload. Engagement, which is related to information gathering, visual scanning and sustained attention (Berka et al., 2007), has been found to have positive impacts on learning (e.g. McQuiggan et al., 2007; Woolf et al., 2007), while the opposite is true of distraction (Froese, 2012). Distraction occurs when attention is withdrawn from processing information necessary to complete the primary task (Strayer et al., 2011), thus leading to decreased resources being focused appropriately to complete a task. Therefore, classifiers of engagement and distraction can be used to infer appropriate attention allocation. High workload, similar to distraction, has shown a negative impact on learning (Gonzalez, 2005). All three states (engagement, distraction, workload) have been previously measured with physiological data such as electroencephalography (EEG) and pupilometry metrics (e.g. Berka, 2007; Ahlstrom and Friedman-Bern, 2006).

For the current study, five low-cost sensors were selected based on the following requirements: 1) collects sensor data shown to be correlated with at least one of the six states, 2) provides ability to collect and access data in real-time, 3) minimally intrusive, and 4) costs less than $500. The sensors included in the study were a motion detector and chair with pressure sensors to capture postural sway, a heart rate monitor, an EEG headset with a single electrode located on the forehead above the left eye (the F_{p1} electrode position), and an eye tracker to capture pupilometry. A full review of the theoretical justification for choosing the affective and cognitive states and the low-cost sensors can be found in Carroll et al. (2011).

EXPERIMENTAL APPROACH

Two experiments were completed to compare low-cost physiological sensor output to validated benchmark measures of affective and cognitive states. The intention of Experiment 1 was to develop classifiers of affective and cognitive states with civilians. Experiment 2 would then utilize these classifiers to validate each with a more relevant population (United States Military Academy Cadets). Both experiments were within-subjects repeated-measures designs in which all participants performed three types of tasks: 1) a visual vigilance task, 2) video clip observation, and 3) Virtual Battlespace 2 (VBS2) scenario completion to induce variations in affective and cognitive state.

The lessons learned from Experiment 1 resulted in redesign of data collection procedures and small changes to the experimental setup for Experiment 2. The purpose of Experiment 2 then became the same as Experiment 1 – to create classifiers of targeted affective and cognitive states with a cleaner set of data (and therefore with a higher possibility of developing generalizable classifiers of affective and cognitive state) and a more relevant population.

EXPERIMENT 1

The objective of Experiment 1 was to evaluate the ability to create real-time classifiers of targeted affective and cognitive states utilizing data from low cost sensors. Validated stimuli to induce each targeted state were presented to participants while data was collected from both low-cost sensors and validated benchmark measures for comparison. By relating data resulting from the low-cost physiological sensors against validated benchmark measures, using logistic regression models, the effectiveness of low-cost sensors at detecting target affective and cognitive states was assessed. Further, development of basic classifiers was attempted based on inputs from the low-cost sensors to model trainee affective and cognitive states. Logistic regression classification was the preferred classification tool for this study, given that logistic regression models are simple, easy to interpret, and well suited to learn relationships between variables expected to be correlated.

Hypotheses

The hypotheses for the experiment were as follows:

1. Good classification models can be created for affective states using data from a series of low-cost sensors and logistic regression techniques.
2. Good classification models can be created for cognitive states using data from a series of low-cost sensors and logistic regression techniques.

Method

Participants

Twenty-five people ranging in age from 19-34 (mean = 25) years participated in the study. Fifteen were male, ten were female, and all but 1 were civilians (the non-civilian had only three months of military experience).

Tasks

Visual Vigilance Task: This task was a three-minute vigilance task in which the user had to press the space key every two seconds with a visual reminder. **Video Clip Observation:** Participants viewed three video clips validated to induce various affective states: 1) a 65-second excerpt from the Warner Brothers film *All the President's Men*, previously validated to induce a neutral affective state, 2) a 236-second excerpt from the 20th Century Fox movie *My Bodyguard*, previously validated to induce anger, and 3) a 208-second excerpt from the Falcon Films movie *Halloween*, previously validated to induce fear (Hewig et al., 2005). **VBS2 Scenarios:** In the VBS2 simulation scenarios, the overall goal of the participant was to search and eliminate enemy threats in an urban environment (e.g., a building or street). To accomplish this, the participant was required to enter a building and move rapidly along a hallway while covering the entire area with their weapon to maintain security. As people were encountered, they were to be quickly evaluated and engaged if hostile (anyone holding a weapon was considered hostile). As participants crossed the threshold of a room doorway that had not been cleared, they performed immediate target engagement of any enemies detected. Each of the four VBS2 scenarios contained an Emotion Induction Technique (EIT; see Table 1), designed to induce either anger or fear in participants, as well as high and low states of distraction, engagement, and workload.

Table 1. Descriptions of Emotion induction Techniques (EITs) within scenarios.

Scen-ario	EIT	Presentation time	Details	Expected Induced States
1	Limiting visual perception (fog)	Throughout scenario	Visibility reduced with dense fog; impaired ability to see enemies	Fear, anger, workload
2	Increasing enemies	Throughout scenario	6-7 enemies per room as opposed to 1-3 in other scenarios	Fear, anger, workload, engagement
3	Annoying sound	First 45 seconds of scenario	Car alarm sound playing	Anger, workload, distraction
4	Equipment malfunction	When participant reaches a physical marker in simulated room	Weapon malfunctions during room clearing	Anger, fear, workload, distraction

Apparatus

Participants used a standard flat screen monitor, keyboard, and mouse throughout the experiment. **Ground-truth affective states** were measured using EmoPro™, an electronic emotional profiling tool that has been validated to accurately measure participants' emotions (Champney and Stanney, 2007). EmoPro™ is a subjective assessment tool by which individuals indicate which emotions they felt during an experience by selecting emoticons representing a number of distinct emotional expressions. Each emoticon is designed to represent one particular emotion by utilizing human expression cues. **Ground-truth cognitive state** was assessed using ABM's B-Alert™ X-10 EEG headset. The associated B-Alert™ analysis software includes EEG indices of workload, engagement, distraction, and drowsiness (Johnson, et al., 2011).

A variety of **non-invasive, low-cost sensors** were used in the study. The NeuroSky MindSet EEG headset collects EEG data from a single-point dry electrode that sits on the forehead above the left eye, in the F_{p1} electrode position. The associated MindSet Research Tools software provides data on a user's Delta, Theta, Alpha, Beta, and Gamma brainwave band power levels, as well as classifiers of Attention and Meditation. The band power levels are output as follows: Delta: 1-3Hz; Theta: 4-7Hz; Alpha1: 8-9Hz; Alpha2: 10-12Hz; Beta1: 13-17Hz; Beta2: 18-30Hz; Gamma1: 31-40Hz; Gamma2: 41-50Hz. The Zephyr HxM™ BT heart rate sensor comes in the form of a strap that is worn around the chest against the skin. It allows for real-time collection of data on a personal computer through a Bluetooth connection. Data collection includes heart rate in beats per minute. The Vernier Go!Motion motion detector uses ultrasound technology to collect position, velocity, and acceleration data of moving objects in real-time through a USB port. When it is placed between a computer and the user, it can detect changes in posture when the user leans forward or sits back in the chair. The chair participants sat in contained eight Phidget pressure sensors (four in the seat, four in the back) to collect pressure data in real time through a USB port, allowing for real-time determination of changes in posture. Finally, a low-cost eye tracker was developed in house for this effort. The hardware of the eye-tracker was composed of a Thorlab DCC1545M monochrome camera with a TVR0614 ½" C Mount 6-15mm F1.4 manual iris lens and an Opteka HD2 37mm R72 720 nm infrared X-Ray IR filter, as well as two IR010 Night Vision IR lights. This hardware was used with the ITU Gaze Tracker open source gaze tracking application to collect pupilometry data. All sensors were sampled synchronously at a common rate of 40 Hz.

In addition to the physiological and behavioral data collected by the sensors, a variety of surveys were administered, including: 1) a demographics questionnaire, 2) the Computer Game Immersion Questionnaire, which determines a participant's tendency to become immersed when playing a computer game, 3) the Life Orientation Test – Revised (LOT-R), which measures trait optimism/pessimism (Scheier, Carver, & Bridges, 1994), 4) the Self-Assessment Manikin (SAM; Lang, 1985), which measures mood in terms of pleasure, arousal, and dominance, and 5) the Neuroticism-Extroversion-Openness Personality Inventory (NEO-PI), a personality questionnaire based on the Big Five personality dimensions (Schinka et al., 1997).

Procedure

Upon arrival, participants received a brief overview of the study and were asked to complete informed consent and surveys. Participants then donned the ABM headset, and filled out the paper-based questionnaires for demographics, immersion, optimism/pessimism, mood, and personality. Next, participants completed the EEG baseline task, and then donned the Zephyr heart rate sensor around their chest, under their shirt. The NeuroSky EEG headset was then placed on their head. The participants sat in the pressure sensor chair in front of a display, motion detector, and eye tracker, and completed a calibration session with the eye tracking system. Adjustments were made to all sensors as needed until continuous data collection was attained.

Once all sensors were in place and successfully collecting data, students performed the series of tasks outlined above to induce variations in cognitive and affective state. First, participants performed a three-minute vigilance task on a personal computer, which consisted of pressing the space bar every time a red circle appeared on the screen. Participants completed an EmoPro™ evaluation just before and just after this task. Next, participants observed three video clips, completing an EmoPro™ evaluation just after each video clip. Next, experimenters described the VBS2 task in detail and had participants go through training to familiarize them with how to interface with the software. Participants were then asked to complete a trial scenario to gain an understanding of what would be expected of them during the experimental task. Next, participants completed a total of four scenarios, all with EITs and with 3-6 critical events per scenario. Following each critical event within a scenario, participants were prompted to complete an EmoPro™ evaluation to indicate their emotional state during the event. Upon completion of the VBS2 scenarios, participants received a short debriefing.

Data Analysis

Data from the eye-tracking and the chair pressure sensors were not considered during analysis due to reliability issues. The eye tracking data (i.e., pupilometry) was confounded by participant movement (stationary eye tracker could not distinguish differences between changes in pupil and changes in distance from sensor) and lighting adjustments (lights turned off during movie clips to increase engagement). The chair pressure sensor data showed extremely low levels of variability, and therefore provided little opportunity to impact any model. It was later determined that the chair sensors tended to detach from their original locations after some use, and thus provided unreliable data.

The remaining low-cost sensor data were aggregated to obtain averages of each metric on a second-by-second basis. To account for inter-individual differences, the heart rate sensor data from each participant was normalized by subtracting the average of resting heart rate captured during the vigilance task. Subsequently, for each metric, a rate of change variable was created to represent the difference between consecutive values in time. The rate of change of the variables was expected to be relevant for modeling the targeted affective and cognitive states.

In the case of data for affective state classification, considering the multiple one-second observations for each event as independent vectors did not provide consistent patterns for any affective attribute. Subsequently, an approach similar to that followed in Picard et al. (2001) was implemented, in which a single vector of aggregated data was obtained from the one-second observations across each event. Three aggregated attributes were created for each sensor for each event: 1) the average of the original variables (Alpha1, Alpha2, etc), 2) the average of the corresponding rate of change variables (Alpha1Diff, Alpha2Diff), and 3) the standard deviation of the original variables (Alpha1Dev, Alpha2Dev, etc). This second level of aggregation provided a single description vector for each event, and was expected to diminish the negative impact of any abnormal sensor data.

The data analysis was conducted using the free statistical language R (http://www.r-project.org/). Based on the correlations between some of the sensors and the different states previously reported in literature, the logistic regression model was selected. Each ground-truth state was converted into a binary response variable by applying a threshold-based procedure that removed the observations associated with state values lying near the corresponding middle point. For EmoPro^{TM} affective self-report values, which ranged from 0 to 5 (0 being absence of the emotion, 5 being intense feeling of the emotion), observations corresponding to values 1 and 2 were disregarded. For cognitive values, which took values in the interval [0,1], observations with state values between 0.3 and 0.7 were also removed from the training data.

In order to assess the performance of the corresponding logistic regression classifiers, three runs of a 10-fold cross-validation procedure were executed for each model, and their corresponding Receiver Operating Characteristic (ROC) curves were plotted. The overall quality of each classifier was assessed by averaging the areas under its ROC curves (AUC values; Fawcett, 2006). bmuht fo elur a sA, excellent classifiers are those having AUC values between 0.9 and 1. Classifiers with AUC values from 0.8 to 0.9 are typically considered good, and those having AUC values from 0.7 to 0.8 are considered fair.

Results

The forward/backward stepwise logistic regression models corresponding to the three affective states (data taken from all participants) were significant at the 0.001 level, and multiple sensors significantly contributed to each model (**Error! Reference source not found.**). However, these models behaved almost like random classifiers when evaluated on observations that were not used for training; i.e., their corresponding AUC values were very near to 0.5. This disagreement between statistical significance of a model and its classification quality is due to the different goals pursued by classical model fitting and building generalizable classifiers.

It was hypothesized that the difficulty in finding logistic regression models for the affective states could be due to high variability between participants. To explore this possibility, the Partitioning Around Medoids (PAM) robust clustering algorithm (Kaufman and Rousseeuw, 1987) was applied to the first five principal components of the data from the demographics questionnaires (we selected the minimum number of principal components giving us a cumulative proportion of variance greater than 70%). PAM effectively separated the group of participants into two subgroups. Only the data set for Anger had enough variability within those subgroups to allow for a separate logistic regression analysis on each of them. A significant logistic regression model was obtained for one of the two subgroups. The model, described in Table 3, included only heart rate, which was negatively correlated to anger. It showed a noticeable deviation from a random classifier when validated using 10-fold cross-validation (Figure 1 shows the corresponding ROC curves). Its average AUC value was 0.6792, which is close to what is typically considered a fair classifier (ROC curves with steeper slopes and therefore higher AUC values are desirable). This result suggested that heart rate obtained from a low-cost sensor could be useful in predicting the presence of anger in a subset of the participants. Furthermore, the fact that clustering facilitated development of this promising result implied that a much more homogeneous group of participants, or a personalized approach, could lead to the development of successful models using low-cost sensors.

The forward/backward stepwise logistic regression models corresponding to each cognitive state for each participant were also statistically significant, but the significant attributes and the signs of the corresponding coefficients were not consistent across for any two participants across all participants. As expected from these results, a 10-fold cross-validation of stepwise logistic regression models obtained from all participants combined resulted in very poor classification performance. As such, it was not possible to obtain good logistic regression classifiers capable of modeling the data from all participants, either for affective or cognitive states.

LESSONS LEARNED FROM EXPERIMENT 1

The results from Experiment 1 yielded little in the way of generalizable classifiers, however much knowledge was gleaned about working with the low-cost sensors in an experimental setting.

Some of the lessons learned stemmed from the amount of components in the experimental testbed, including both hardware and software. In Experiment 1, the stationary eye tracker was unreliable in tracking pupilometry given free fore/aft movement of the participant (which was captured via posture sensors). A low-cost head-mounted eye tracker was developed and utilized in Experiment 2 to eliminate this issue. Further, with the heart rate monitor in Experiment 1, the chest strap did not have good conductance, resulting in heart rate values of zero. It was determined that this was due to a lack of moisture on the sensor (as it is designed to be used during workouts that result in perspiration). Thus, during Experiment 2, water was used to wet the pads of the chest strap that sat against the skin to improve the connection and minimize data loss.

Also during Experiment 1, heart rate during the vigilance task was used to normalize participant's data. To improve the baseline data capture, a task was included in Experiment 2 in which the participant just stared at a blank screen for 30 seconds during which they were asked to just relax and try not to move. They were told explicitly that the purpose of the task was just to collect baseline data from the sensors.

Another challenge was the time required to ensure good data collection from the low-cost sensors. Due to continued need to troubleshoot sensors throughout Experiment 1 (sometimes between each task), the experiment ran longer than expected, and exceeded the strict time constraints imposed in Experiment 2 due to the scheduling conflicts associated with West Point Cadets. To address this issue, the stimuli found in the first study to induce the greatest variability in both cognitive and affective states, the VBS2 scenarios, were moved in the script to the beginning of the experimental session, right after the baseline task. Therefore, if time ran out, only the movies and/or vigilance task would not be completed.

Another challenge was in trying to make the testbed immersive in order to induce the targeted affective and cognitive states, while still meeting the assessment requirements. Due to the lack of validated, objective, unobtrusive methods to assess affective state, it was necessary to stop the participant at certain intervals to obtain self-report affective state assessment using EmoPro™. During the simulation scenarios, this was done by having the scenario automatically pause for the participant to complete EmoPro™ after each predefined scenario event. Each time the scenario paused, the immersion was broken, with the risk that emotions may not be felt as intensely throughout the scenario as they could be if scenario interaction went uninterrupted. On the other hand, it is critical that enough events are completed to get the maximum amount of data. It can be difficult to balance these competing requirements. To compensate for some of the loss of immersion, Experiment 2 incorporated small changes to the VBS2 scenarios in order to try to increase emotional intensity within each event, such as moving the enemy into a position within a room where they could not be seen from the doorway. Therefore, a number of changes to the experimental protocol were made as outlined above to reduce data variability and data loss in Experiment 2.

EXPERIMENT 2

Method

Experiment 2 was conducted at the United States Military Academy. Due to the strict time constraints of Cadets, the study was run in two parts: Part 1 took approximately one hour, and included performing the baseline procedure for EEG and completing all questionnaires; Part 2 took approximately two hours and consisted of donning the neurophysiological sensors, and then performing the baseline task, VBS2 scenarios, vigilance task, and observing the three movie clips.

Participants

A power analysis similar to that done in the first experiment was conducted, and it was determined that a minimum of 18 participant was necessary. Twenty participants completed the entire experiment, all West Point Cadets with active duty experience (including their time at West Point) ranging from 9-44 months (average of 15.45 months). Majority of participants were first-year cadets enrolled in the Behavioral Sciences and Leadership (BS&L) department's General Psychology (PL100) course. Ages ranged from 18-23 years, with an average of 19 years.

Data Analysis

Due to the improvements to the experimental setup, reliability and quality of the sensor data were greatly improved and data from all sensors was able to be included in this analysis. The data from the different sensors were processed in the same manner as for Experiment 1, except that for the chair sensors and pupil diameter, for which only their averages were included in the training datasets (i.e., no standard deviation and rate of change were calculated for them). This was due to low data variability of the chair sensors in short periods of time and the need to remove about 15% of pupilometry data that was clearly out of range.

Results

Hotelling T^2 tests were run on the datasets corresponding to the affective states to determine whether there were statistically significant differences between the means of sensor data corresponding to the presence of emotion versus the absence of it. The tests first failed to run because of multi-collinearity issues. Multi-collinearity diagnostics based on the inflation of variances (Fox and Monette, 1992) detected that many EEG bands from the Neurosky MindSet were highly correlated with other attributes in the data. Those attributes were removed from the input to the T^2 tests. We were able to reject the hypothesis of equal means in the cases of Boredom and Fear at 0.001 significance level. However, that null hypothesis could not be rejected for Anger ($p = 0.3645$).

The analysis then focused on finding accurate logistic regression classifiers for the three affective states. Additionally, logistic model trees (LMTs) were explored as a classification appraoch. The LMT algorithm produces decision trees that contain logistic regression functions at their leaves (Landwehr et al., 2005); with the simplest LMT classifier being a single node containing a logistic regression model per class. LMT provides the capability of dealing with non-linear relationships in the data while still offering a model that is easy to interpret. The LMT algorithm from the "RWeka" R package was employed to obtain the LMT models.

As in the first experiment, the classification accuracy of the models was evaluated through the average AUC values from 10-fold cross-validation, although this time 10 runs of the cross-validation procedure were executed. The standard deviations of the AUC values were also calculated. None of the logistic regression models (on all variables and also using stepwise regression) showed a good or even fair generalization capability. However, the LMT models obtained for Boredom and Fear showed good generalization capability. The model for Boredom showed an average AUC value equal to 0.79 with 0.008 standard deviation (Figure 1). The model for Fear had an average AUC value equal to 0.83 with 0.012 standard deviation (Figure 2).

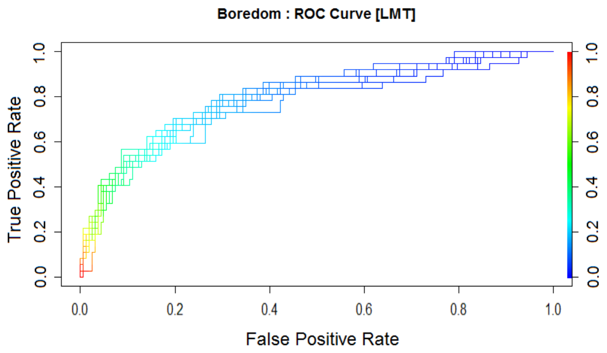

Figure 1. ROC curves for LMT model of Boredom.

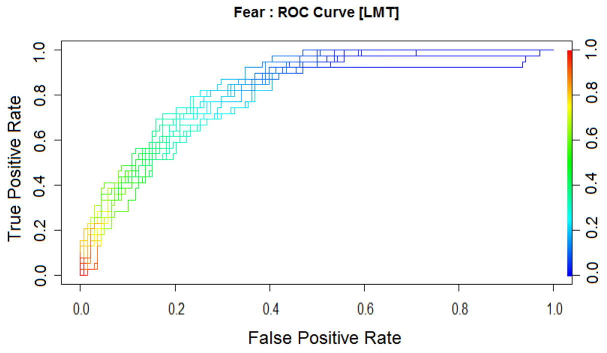

Figure 2. ROC curves for LMT model of Fear.

Curiously, each resulting LMT model consisted of a single logistic regression model, suggesting that the LMT algorithm is particularly efficient at finding logistic regression models that favor generalization over statistical significance, thus avoiding overfitting. This is due to the LogitBoost algorithm (Friedman, Hastie, & Tibshirani, 2000) used in LMT to fit the logistic regression models at the nodes of the tree. Table 2 provides the attributes that were factors in the LMT models for the affective states and the associated sensors. These results suggest that data obtained from these low-cost sensors could be useful to predict the presence of boredom and fear in learners. Statistical significance values are not provided for the LMT logistic regression models; contrary to classic logistic regression, statistical significance is not important to the LMT learning algorithm.

Table 2. Factors in Affective State LMT Models

Affective State	Low-Cost Sensor	Attribute	Coeff.Sign
Boredom	EEG	(Intercept)	-
		Alpha2	+
		Gamma	-
		Gamma2Diff	-
		BetaDiff	-
		BetaDev	+
		Beta2Diff	+
		AttentionDev	+
	Heart Rate Monitor	HeartDiff	-
		HeartDev	+
	Distance Sensor	MotionDev	+
Fear	EEG	(Intercept)	-
		Gamma2Dev	-
		Beta2Dev	+
		Delta	-
		Attention	-
	Distance Sensor	Motion	+
	Chair Seat Sensors	ChairSensor6	-
		ChairSensor7	+

Based on the results from the first experiment, it was expected that a classifier could be obtained for the Anger state. Unfortunately, it was not possible to obtain even a fair classifier for that state using the two models considered. This result agrees with the Hotelling T^2 test results. It was hypothesized that the lack of significance for difference of means in the case of Anger could be due to the presence of outlier observations, so the multivariate robust outlier detection method proposed in Filzmoser et al. (2005) was applied to the Anger data set, and observations labeled as outliers were removed. The T^2 test was repeated on the reduced Anger data set. The p-value obtained was equal to 0.0662, which still did not allow for rejection of the null hypothesis at 0.05 or lower significance level. A logistic regression model was obtained from the reduced Anger data set using the LMT algorithm. A 10-fold cross-validation of that model showed that it was an almost random classifier (average AUC = 0.58). As a final attempt to obtain a good classifier for Anger, a clustering approach similar to the one used in the first experiment was applied. A PAM clustering of the demographic data using the first 5 principal components (which gave us up to 75% of variance) found two groups of participants. One of the groups consisted only of three participants, which did not provide enough data for fitting a model to its corresponding dataset. In the case of the other group, a stepwise logistic regression model that included the Delta and BetaDiff attributes turned out to be significant at 0.01 alpha level. However, its corresponding average AUC value was 0.48, which denotes a random classifier. Cross-validation of the LMT model on this data also gave a very poor average AUC value: 0.60.

Only those participants that completed the whole experiment and had a good benchmark EEG baseline and data files (10 participants in total) were considered for the analysis of the cognitive states. This was not a limitation for the analysis, given that ground-truth values for cognitive classifiers were collected every second for about 2 hours for each participant during the experiment. Consequently, a large amount of training data was available from those 10 participants. Following an approach similar to that used for the affective states, the logistic regression models were cross-validated on the cognitive data sets from all 10 participants combined. None of the logistic regression models showed a good or even fair generalization capability. However, exploratory 10-fold cross-validations of the LMT models on all attributes showed good generalization capability.

The LMT models obtained from all the training data on all the attributes consisted of highly complex decision trees. Too much complexity in machine learning models is typically associated to overfitting the training data, which might lead to poor classification performance on new data. The complexity of the LMT models pointed to the need of finding subsets of relevant features for each cognitive state, in order to obtain simpler LMT models, while still keeping similar or better generalization capability (AUC values). The feature selection techniques provided by the Boruta R package (Kursa and Rudnicki, 2010) and the standardized coefficients from Linear Discriminant Analysis (LDA) (Rencher and Scott, 1990) were therefore employed. The mean and standard deviation of AUC values were obtained from 5 cross-validation runs on up to the first 10 variables selected by Boruta and LDA, separately. Based on these results from the feature selection techniques, LMT models were trained on several combinations of the most relevant variables and models achieving both good AUC values and low model complexity were selected. Table 3 provides the attributes that were factors in the LMT models for the cognitive states and the associated sensors. The average AUC values and their standard deviations were calculated through 10 runs of 10-fold cross-validations. The average AUC for Distraction was 0.81, with 0.010 standard deviation. The average AUC for Engagement was 0.80, with 0.004 standard deviation. The average AUC for Workload was 0.82, with 0.008 standard deviation.

Table 3. Factors in Cognitive State LMT Models

Cognitive State	Low-Cost Sensor	Attribute
Distraction	Heart Rate Monitor	Heart Rate
	Chair Seat Sensors	All 4 Seat Sensors
Engagement	Heart Rate Monitor	Heart Rate
	Chair Seat Sensors	All 4 Seat Sensors
	Distance Sensor	Motion
Workload	Distance Sensor	Motion
	Chair Seat Sensors	All 4 Seat Sensors

CONCLUSIONS AND FUTURE RESEARCH

These experiments evaluated low-cost sensors for utilization in classifying cognitive and affective states, with results providing preliminary evidence of their utility in computer-based training environments. Good classification models based on data from low-cost sensors have been developed for the affective states of fear and boredom, and the three cognitive states considered in this study.

A classifier for Anger could not be learned by the two models considered in this study. Several techniques suggested that our dataset for Anger did not provide enough differences between the presence and the absence of that affective state: (1) Hotelling T^2 tests were not able to reject the equal means hypothesis; (2) removing outliers did not facilitate learning a good classifier; and (3) building a classifier from a subset of participants that shared similar demographics did not provide a good classification accuracy.

Based on the models obtained, some of our hypotheses were partially or completely met. Regarding the affective states, the final logistic regression models obtained from LMT showed that heart rate and posture measures were factors in the model of Boredom, and posture sensors were factors in the Fear model. As expected, posture sensors were factors in the engagement model, but unexpectedly also contributed to the distraction and workload models, and heart rate was an unanticipated factor in the distraction and engagement models. However, low-cost EEG attributes factored into both affective state models, but none of the cognitive state models. Also, pupil diameter did not play a factor in any of the models. These results were surprising, given the amount of literature support for using EEG and pupilometry metrics to measure cognitive states. This may lend evidence to the poor reliability of the low-cost sensors. The fact that the NeuroSky EEG had a single electrode and that the eye-tracking sensor was noisy could be reasons for the lack of correlation. It is important to recall that about 15% of the pupil diameter data had to be removed and more noisy observations might have remained in the datasets.

Given that the cutoff values to be used with the models depend not only on classification accuracy but also on the risks associated to each type of classification error (false positives and false negatives), conclusive cutoff values are not provided here. However, assuming the same risk for both types of classification error, the best cutoff values for each model seem to be around the following numbers: Boredom: 0.2; Fear: 0.2; Distraction: 0.2; Engagement: 0.3; Workload: 0.8.

Future steps will include integrating the sensors and models into the Generalized Intelligent Framework for Tutors (GIFT; Sottilare, et al., 2011), a domain-agnostic ITS architecture. To complete integration, sensor-specific interfaces will need to be developed to capture raw sensor data for processing and eventual classification of either a cognitive or affective state with a high degree of accuracy. Once candidate sensors have been integrated into the GIFT (Sottilare, et al., 2011) a series of assessments can begin. As noted in the lessons-learned section of this paper, there may be significant incompatibilities between sensors in a given sensor-state group, and the most accurate sensors may not be the most practical (e.g., a sensor with low usability for a given task). So while this particular research has been instrumental in narrowing the field of sensors, significant research lies ahead to determine the smallest compatible set of sensors to predict each cognitive and affective state.

Overall, this research provides evidence to support the theory that, with the help of low-cost sensors, ITSs can begin to rival the effectiveness of human tutors by diagnosing affective and cognitive states that contribute to a decrease in readiness to learn. Future work must also determine appropriate learning strategies to implement during periods of low readiness to learn and how to implement them. Furthermore, as sensor technology improves, sensors will be less invasive, will cost less, and will become more accurate. Subsequently, classifiers of more affective states can be developed, and those already developed will become more reliable.

ACKNOWLEDGEMENTS

This research was sponsored by the U.S. Army Research Laboratory - Human Research and Engineering Directorate under contract W91CRB-10-C-0327 and the U.S. Office of Naval Research under contract N00014-08-C-0535.

REFERENCES

Ahlstrom, U., & Friedman-Bern, F.J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics,36*(7), 623-636.

Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Zivkovic, V.T., Olmstead, R.E., Tremoulet, P.D., Craven, P.L. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation Space and Environmental Medicine*,*78*(5, Suppl.), B231-B244.

Burleson, W. & Picard, R. (2004). Affective Agents: Sustaining Motivation to Learn Through Failure and a State of “Stuck".*Paper presented at the Workshop on Social and Emotional Intelligence in Learning Environments at the 7th International Conference on Intelligent Tutoring Systems*, Maceio, Alagoas, Brazil.

Calvo, R. A., & D'Mello, S. (2010). Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing, 1*(1), 18-37.

Carroll, M., Kokini, C., Champney, R., Fuchs, S., Sottilare, R., & Goldberg, B. (2011). Modeling trainee affective and cognitive state using low cost sensors. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) Annual Meeting*. Orlando, FL.

Champney, R.K., & Stanney, K.M. (2007). Using Emotions in Usability. Proceedings of the Human Factors and Ergonomics Society 51th Annual Meeting. Baltimore, Maryland, October 1-5.

D’Mello, S. K., Taylor, R., & Graesser, A. C. (2007). Monitoring Affective Trajectories during Complex Learning. In D. S. McNamara and J. G. Trafton (Eds.). *In Proceedings of the 29th Annual Cognitive Science Society*, 203-208. Cognitive Science Society, Austin.

Fawcett. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861–874.

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association, 87*, 178–183.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics, 38*(2), 337–374

Froese, A. D., Carpenter, C. N., Inman, D. A., Schooley, J. R., Barnes, R. B., Brecht, P. W., & Chacon, J. D. (2012). Effects of classroom cell phone use on expected and actual learning. *College Student Journal*,*46*(2), 323-332.

Gonzalez, C. (2005). The relationship between task workload and cognitive abilities in dynamic decision making. *Human Factors*, *47*(1), 92–101.

Graesser, A., Chipman, P., Haynes, B., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, *48*(4), 612-618.

Graesser, A.C. & D’Mello, S. (2012). Emotions during the learning of difficult material. In B. H. Ross (Ed.), *Psychology of Learning and Motivation*, *57*, 183-225.Academic Press.

Hewig, J., Hagemann, D., Seifert, J., Gollwitzer, M., Naumann, E., & Bartussek, D. (2005). A revised film set for the induction of basic emotions. *Cognition and Emotion*, *19*(7), 1095-1109.

Johnson, R. R., Popovic, D. P., Olmstead, R. E., Stikic, M., Levendowski, D. J., & Berka, C. (2011). Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model. *Biological Psychology, 87*, 241-250.

Kaufman, L., & Rousseeuw, P.J. (1987). Clustering by Means of Medoids. In Y. Dodge (Eds.).*Statistical Data Analysis Based on the L_1–Norm and Related Methods*. North-Holland, 405–416.

Kursa, M.B.,& Rudnicki, W.R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software, 36*(11), 1-13. http://www.jstatsoft.org/v36/i11/

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic Model Trees. *Machine Learning*, *59*(1), 161-205.

Lang, P.J. (1985). *The Cognitive Psychophysiology of Emotion: Anxiety and the Anxiety Disorders.* Hillsdale, NJ; Lawrence Erlbaum.

Lisetti, C.L., & Nasoz, F. (2004). Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals. *Journal on Applied Signal Processing*, *11*, 1672-1687.

Lepper, M., & Woolverton, M. (2002). The Wisdom of Practice: Lessons Learned from the Study of Highly Effective Tutors. In J. Aronson (Ed). *Improving academic achievement: impact of psychological factors on education*, 135-158. New York: Academic Press.

McQuiggan, S., Lee, S., & Lester, J. (2007). Early prediction of student frustration. *Affective Computing and Intelligent Interaction*, 698-709.

Picard, R.W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(10), 1175-1191.

Rencher, A. C., & Scott, D. T. (1990). Assessing the Contribution of Individual Variables Following Rejection of a Multivariate Hypothesis. *Communications in Statistics: Simulation and Computation*, *19*(2), 535–553.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A re-evaluation of the Life Orientation Test. *Journal of Personality and Social Psychology, 67*, 1063-1078.

Schinka, J. A., Kinder, B. N., & Kremer, T. (1997). Research Validity Scales for the NEO--PI--R: Development and Initial Validation. *Journal of Personality Assessment*, *68*(1), 127-138.

Sottilare, R., Holden, H., Brawner, K., & Goldberg, B. (2011). Challenges and Emerging Concepts in the Development of Adaptive, Computer-based Tutoring Systems for Team Training. *Interservice/Industry Training Systems & Education Conference*, Orlando, Florida, December 2011 .

Stevens, R. H., Galloway, T., & Berka, C. (2007). Integrating Innovative Neuro-Educational Technologies (I-Net) into K-12 Science Classrooms. In D. Schmorrow & L. Reeves (Eds.), *Foundations of Augmented Cognition - Proceedings of the Third International Conference*, *HCI International 2007. LNCS,4565*, 47-56. Berlin: Springer.

Stottler, D., Harmon, N., &Michalak, P. (2001). Transitioning An ITS Developed For Schoolhouse Use To The Fleet-TAO ITS, A Case Study.*Paper presented at the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*, Orlando, FL.

Strayer, D. L., Watson, J. M., & Drews, F. A.(2011). Cognitive Distraction While Multitasking in the Automobile. In B. Ross, (Ed).*The Psychology of Learning and Motivation, 54*, 29-58. Burlington: Academic Press.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., & Wintersgill, M. (2005).The Andes Physics Tutoring System: Five Years of Evaluations. *Paper presented at the International Conference on Artificial Intelligence in Education, Amsterdam.*

Woolf, B., Burleson, W., &Arroyo, I. (2007). Emotional intelligence for computer tutors. *Supplementary Proceedings of the 13th International Conference of Artificial Intelligence in Education*, 6-15.

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: recognizing and responding to student affect, *International Journal of Learning Technology*, *4*(3/4), 129–164.