Efficacy of Measuring Engagement during Computer-Based Training with Low-Cost Electroencephalogram (EEG) Sensor Outputs

Benjamin Goldberg[a], Keith W. Brawner[a], and Heather K. Holden[a]
[a)] US Army Research Laboratory-Human Research & Engineering Directorate-Simulation and Training Technology Center- Learning in Intelligent Tutoring Environments (LITE) Lab

The potential of Intelligent Tutoring Systems (ITS) to influence learning may be greatly enhanced by the system's ability to accurately assess a student's cognitive state in real-time. For this to happen, interactions with, and reactions to, training content must be collected and assessed; data is then used to inform instructional adaptation within the system. Validated sensors are available for this purpose and have been shown to correlate with cognitive and affective states linked to learning. However, sensors used to inform student models are often expensive and impractical for wide-range use. In this paper the authors present a study evaluating the efficacy of using Emotiv's Electroencephalogram (EEG) Affectiv Suite outputs to inform an ITS student model. In this experiment, seventy-three participants interacted with the Cultural Meeting Trainer (CMT), a web-based cultural negotiation trainer, while Emotiv's engagement, short-term excitement, and long-term excitement metrics were indexed and logged. Our analysis assesses the quality of Emotiv metrics across one well-defined and two ill-defined scenarios. Results show consistent outputs across tasks and support further examination into the Emotiv's ability to accurately track cognitive state in a learning environment.

INTRODUCTION

The functionality of Intelligent Tutoring Systems (ITS) has historically centered around theories of education that view thinking and learning as cognitive processes bound to a problem space (Woolf et al., 2009). While ITSs designed under this notion have been found to be effective in well-defined academic domains (Koedinger, Anderson, Hadley, & Mark, 1997; Woolf, 2009), there are significant limitations to the current state of the art, with a large emphasis on examining the relationship between student affect and learning outcomes (Shute, 2007). Tracking states found to impact learning and retention will result in systems that recognize affect and respond with interventions that encourage effort, lessen humiliation, and provide support and motivation for further interaction. A state of interest linked with information gathering, visual scanning, and periods of sustained attention is engagement (Berka et al., 2007). Affective elements such as engagement and motivation are linked with cognition in that they guide memory and decision making processes (Norman, 1981). They have also been shown to have a significant influence on learning outcomes (Craig, Graesser, Sullins, & Gholson, 2004; Woolf et al., 2009).

Though these elements have been found to impact learning, ITS capabilities are bound to the available data streams they can access pertaining to an individual user. This requires identifying student modeling tools and methods for collecting affect relevant information that can be used to predict and track affective states in real-time (Sottilare, Goldberg, & Durlach, 2011). This data is used in conjunction with performance metrics to determine when and how to adapt instructional content. Detection of engagement, how it changes over time, and how it changes in response to stimulus can be used to select instructional strategies that maximize performance on the individual level.

While human tutors are apt to recognize student engagement from visual and auditory cues, computer-based tutoring systems must distinguish these states from sensing technologies that monitor physiological and behavioral markers. Electroencephalogram (EEG) is a physiological variable of electrical activity along the scalp, and has been found to correlate with attention, memory and perception (Fabiani, Gratton, & Coles, 2000). Sensors such as the Advanced Brain Monitoring wireless B-Alert EEG system have been used in this context to assess workload and engagement measures (Coyne et al., 2010). However, the cost associated with these sensors, although appropriate for the research setting, is not cost-effective for widespread application. The goal of this research is to assess the efficacy of using a low-cost EEG headset to inform student models of real-time engagement levels. The Emotiv EPOC was selected for this study because it was under $500, with the major limitation that it does not provide raw EEG data streams. The research strategy is to determine the accuracy of their proprietary affective metrics by monitoring sensitivity and trends to variations in stimuli over time. A secondary objective is to identify correlations between Emotiv outputs and self-reported levels of engagement.

METHODS

Subjects

Seventy-nine cadets enrolled at the United States Military Academy (USMA) at West Point were recruited as volunteer subjects for the experiment. USMA cadets were selected because they represent an Army relevant population of future Officers who will potentially interact with ITS integrated training platforms. They also represent an ideal sample for a university student population who lack specific skill sets, which is a key focus for development of such technology. Of the 79 subjects, 59 were male and 20 were female. Participant age ranged between 18 and 25 years of age (M = 19; SD = 1.25) and all were registered at the time in the PL100 General

Psychology course. All subjects reported no previous training in cultural negotiations prior to participation. Of the seventy-nine subjects, six data sets were removed due to signal dropouts, leaving 73 usable sets.

Apparatus

Emotiv. All participants were fitted with the Emotiv EPOC neuro-headset (see Figure 1), a commercial-off-the-shelf Electroencephalogram (EEG) brain-computer interface. The Emotiv is composed of 14 electrodes with locations following the American EEG Standard ("American Electroencephalographic Society Guidelines for Standard Electrode Position Nomenclature," 1991). Electrodes are felt-based with gold-plated contacts and use saline solution for conductance. To maintain low-cost, the proprietary software development kit's Affectiv Suite tool was used as is, avoiding the purchase of additional licenses required for accessing raw data channels.

Figure 1. 14-Channel Emotiv EPOC Neuroheadset

The Affectiv Suite reports three detection states in real-time: Short-Term Excitement (STE), Long-Term Excitement (LTE), and Engagement (ENG). The tool looks for distinct brainwave characteristics that are universal in nature and do not require signature-building or individual baselining (Emotiv Systems). These detection states are the primary dependent variables examined. The purpose of this study is to test the accuracy of these detection states in a training context and assess their applicability for integration within ITS frameworks. As described by Emotiv, the excitement metrics are associated with positive feelings of arousal, and are characterized by physiological responses including pupil dilation, eye widening, and increases in heart rate and muscle tension. Outputs are produced that represent trends for short- and long-term time segments, with increases in physiological arousal resulting in higher detection scores (Emotiv Systems). In comparison, Emotiv defines engagement as the conscious management of attentional resources towards task-relevant stimuli; greater the focus and cognitive workload, higher the output score. This is characterized by increases in beta and attenuated alpha waves, which are both well-known types of EEG wave-forms (Emotiv Systems).

Cultural Meeting Trainer. To test the efficacy of using Emotiv's metrics for informing student models, a testbed was selected allowing tailored authoring of scenario content within a computer-based training platform. This enabled the creation of scenario conditions that range in task complexity and interaction characteristics. The manipulation of these variables is intended to produce differing levels of cognitive load and engagement for assessing Emotiv's detection capabilities. The resulting study was conducted using the Cultural Meeting Trainer (CMT), a web-based flash system prototype applied for cross-cultural interaction training. The CMT (see Figure 2) is based on the U.S. Army's Bilateral Negotiations Trainer, an immersive virtual environment that allows practice and execution of face-to-face negotiations with virtual humans that include cultural models (e.g., Iraqi Culture) (Kim et al., 2009). The CMT is specifically designed to facilitate the training of cross-cultural norms and customs associated with conversation and discussion leading up to negotiation. Subjects interact with CMT characters through a bank of dialog selections used for progressing story paths.

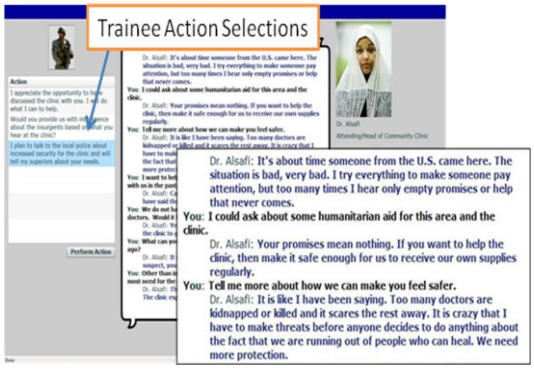

Figure 2. Screenshot of Cultural Meeting Trainer (CMT) Trainee Interface

To assess Emotiv's capacity for tracking cognitive states across tasks of varying complexity, a counterbalanced within-subjects experiment was designed. Three conversational scenarios were developed in the CMT with manipulations to two independent variables: (1) clarity of task execution and (2) presence or absence of character interruptions. Task clarity is based on how well mission objectives are defined. A well-defined task follows a clear set of procedures for achieving mission success, while ill-defined tasks are associated with having ambiguous and vague objectives and comprise multiple approaches for achieving goals. Executing ill-defined tasks requires subjective reasoning and on-the-spot decision making. Variations in task clarity provide the ability to determine if Emotiv metrics produce observable differences between well-defined and ill-defined problem spaces. The second manipulation, presence or absence of interruption, occurs in one of the two ill-defined scenarios. This interaction allows assessment of the effect an interruption in task execution has on Emotiv metrics.

To test these independent variables against Emotiv metrics, three interactive scenarios were developed: Well-Defined No Interruption (WDNI), Ill-Defined No Interruption (IDNI), and Ill-Defined Interruption (IDI). Each scenario is based on the same storyline. The participant is instructed to act the role of a squad leader and is ordered to converse with hospital staff following a nearby insurgency attack. Each scenario is prefaced with a brief description of the interacting character and the objectives associated with the meeting. The WDNI task involves a new in-house physician and requires the participant to maintain casual small-talk while obeying cultural norms for the purpose of building rapport. The IDNI

scenario involves the hospital's lead physician. The goal is to gather information regarding the attack while avoiding U.S. commitments and assurances. The IDI task includes the hospital administrator with objectives for gaining U.S. support among hospital staff and determining current needs to sustain efficient practices. The IDI scenario incorporates an interruption in task progression by having the administrator suddenly speak out of turn. This will allow for determining if a break in the expected flow of interaction will have an effect on engagement and attention levels. Based on these scenario conditions, the following hypotheses were evaluated: $\mathbf{H_1}$: Emotiv will produce reliably different outputs across all metrics when comparing rest with task execution; $\mathbf{H_2}$: The interruption within the IDI scenario will produce a noticeable response in Emotiv metrics reliably across participants; $\mathbf{H_3}$: Emotiv metrics will show reliable differences between conditions; and $\mathbf{H_4}$: Emotiv outputs will correlate with self-reported levels of engagement and mood.

Procedure

Upon arrival participants reviewed a description of the study and signed informed consent. Next, each subject was fitted with an Emotiv EPOC and given a demographics questionnaire. Information pertained to the participant's age and education level, and prior experience in intercultural negotiations. This was followed by an introduction to the CMT interface through an initial conversation with a virtual character. The conversation provided a review of the study and allowed the participant to ask questions about the interface components and the sensor technologies they were wearing. This led into the first of three scenarios with conditions presented in random order across participants. Before each conversation, participants were instructed to relax during a two-minute window. The intention of this break is to mediate the effects of prior interactions on the Emotiv metrics, and to put subjects in a relaxed state before resuming task execution.

At the end of each scenario, a 2-part self-report instrument was administered for the collection of dependent variables of interest. The first survey presented fourteen "engagement-specific" ($\alpha = .89$) items pulled from the Independent Television Commission-Sense of Presence Inventory (ITC-SOPI). This bank of questions was selected because presence has been found to be highly correlated with attention signals and reasonably correlated with engagement in virtual environments (Lombard, Ditton, & Weinstein, 2009; Tang, Biocca, & Lim, 2004). All 14-items are scored on a 5-point Likert-scale, with the mean providing an overall engagement measure. The second survey presented to each participant post-conversation was the Self-Assessment Manikin (SAM), a validated non-verbal graphic-based instrument used for evaluating Mehrabian's three dimensions of mood: pleasure, arousal, and dominance (Bradley & Lang, 1994). Each dimension is scored on a 9-point Likert-scale, and participants are instructed to mark the point most closely resembling their current state. A vector score is calculated across all three dimensions to produce a mood metric. These variables were collected for identifying correlations between self-report measures and the Emotiv physiological outputs.

RESULTS

In determining Emotiv's efficacy for informing student models, several statistical tests were performed to evaluate its ability for tracking cognitive state over time. For analysis purposes, Emotiv data was segmented and post-processed in the following ways. Across all three outputs (STE, LTE, and ENG), averages were calculated within specified time windows for each rest phase and scenario condition. This enables the ability to track the Affectiv Suite outputs across time and observe differences in measures as a participant interacts with the system. Each scenario was divided into three time windows based on length of execution. A mean for the corresponding time window was calculated and used for comparative evaluations. In conjunction, a single mean for each rest phase was used to compare scenario segments associated with the conversation directly following.

An initial test was run to identify if Emotiv's Affectiv Suite outputs could reliably detect differences between rest states and training task execution. A repeated-measures Analysis of Variance (ANOVA) was conducted within each scenario condition to identify significant differences between associated time segments. This is to observe the trend in output metrics as individuals transition from a rest state into task interaction, and to view the effect time within scenario has on excitement and engagement scores. Results show significant differences across all Emotiv metrics. The following table displays ANOVA results for each condition.

Table 1. ANOVA Results Comparing Windowed Time Segments across Each Scenario Condition and Emotiv Metric

	n	F	df	p-value
WDNI				
Short-Term Excitement (STE)	73	83.060	(1, 72)	<.001
Long-Term Excitement (LTE)	73	94.307	(1, 72)	<.001
Engagement (ENG)	73	68.571	(1, 72)	<.001
IDNI				
Short-Term Excitement (STE)	73	59.512	(1, 72)	<.001
Long-Term Excitement (LTE)	73	92.201	(1, 72)	<.001
Engagement (ENG)	73	53.543	(1, 72)	<.001
IDI				
Short-Term Excitement (STE)	73	58.868	(1, 72)	<.001
Long-Term Excitement (LTE)	73	94.639	(1, 72)	<.001
Engagement (ENG)	73	78.387	(1, 72)	<.001

Post-hoc analysis was conducted on all conditions to compare means of rest and scenario time segments. The first investigation was to identify trends in metrics going from rest phase into the first time segment of a scenario condition. The ENG metric shows to have the greatest difference in mean value between the rest and segment 1 window for all scenarios (ENG-WDNI-Rest [M = .471, SD = .016] vs. ENG-WDNI-Segment1 [M = .622, SD = .008]; ENG-IDNI-Rest [M = .485, SD = .016] vs. ENG-IDNI-Segment1 [M = .604, SD = .009]; and ENG-IDI-Rest [M = .470, SD = .014] vs. ENG-IDI-Segment1 [M = .618, SD = .009]). This is the only significant

difference between time segments for the ENG metric, as ENG values become stable once task execution begins.

For excitement metrics, the only condition to show reliable differences for both the STE and LTE mean values going from Rest through Segment1 was IDI (STE-IDI-Rest [*M* = .556, *SD* = .025] vs. STE-IDI-Segment1 [*M* = .456, *SD* = .021]; and LTE-IDI-Rest [*M* = .573, *SD* = .023] vs. LTE-IDI-Segment1 [*M* = .484, *SD* = .020]). IDNI condition produced reliable differences for only the LTE values when comparing the Rest phase (*M* = .577, *SD* = .021) against Segment1 (*M* = .496, *SD* = .019), and no significant differences were found for WDNI. The largest effect seen in both excitement metrics occurs between Segment1 and Segment2 (see Table 2), with significant differences seen within all scenario conditions.

Table 2. Pairwise Comparison Results for Emotiv Excitement Metrics between Scenario Time Segment 1 and Segment 2

	n	Mean	Standard Deviation
Short-Term Excitement (STE)			
STE-WDNI-Segment1 vs. STE-WDNI Segment 2	73	0.507	0.022
	73	0.340	0.018
STE-IDNI-Segment1 vs. STE-IDNI Segment 2	73	0.479	0.022
	73	0.357	0.019
STE-IDI-Segment1 vs. STE-IDI Segment 2	73	0.456	0.021
	73	0.337	0.017
Long-Term Excitement (LTE)			
LTE-WDNI-Segment1 vs. LTE-WDNI Segment 2	73	0.553	0.020
	73	0.392	0.017
LTE-IDNI-Segment1 vs. LTE-IDNI Segment 2	73	0.496	0.019
	73	0.372	0.019
LTE-IDI-Segment1 vs. LTE-IDI Segment 2	73	0.484	0.020
	73	0.349	0.016

Next, data was arranged for the purpose of testing the effect interruption in task execution has on Emotiv metrics. A mean-difference variable was calculated between Segment3 and Segment1 for both ill-defined scenarios. This approach is based on knowing the interruption occurs within Segment2 while task clarity is controlled for. Observing the difference in measures for the start and completion of each ill-defined scenario is used to determine if the presence of an interruption produces detectable changes in Emotiv outputs. No significant differences were found between the ill-defined conditions. This approach was also used to examine the effect task clarity has on the Emotiv metrics over time within subjects. The same mean difference variable was created for WDNI and compared against both ill-defined scenarios. Running a repeated-measures ANOVA on the mean-differences for all three conditions, the tests of within-subjects contrasts showed differences between scenarios for the two excitement metrics (STE: F (1, 72) = 4.117, p <.05; and LTE: F (1, 72) = 6.813, p <.025). Through pairwise comparison, only LTE showed reliable differences between ill- and well-defined conditions (IDI: [*M* = -.139, *SD* = .021] and IDNI: [*M* = -.125, *SD* = .021] vs WDNI [*M* = -.211, *SD* = .023]).

Analysis was further conducted to observe between-scenario differences of the Emotiv metrics by comparing the same time segments across conditions. This can inform if there are significant differences in metric values between conditions and where in time-on-task these variations are produced. Results from a repeated-measures ANOVA show significant differences across conditions for STE in time segment1, F = 4.509, p<.05; LTE in time segment1, F = 11.975, p<.01; and LTE in Time Segment2, F = 4.416, p<.05. Performing a post-hoc pairwise comparison shows significant differences for STE in Time Segment1 between the IDI (*M* = .456, *SD* = .021) and WDNI (*M* = .507, *SD* = .022) conditions. Significant differences between conditions were also found for LTE Time Segment1 when comparing IDI (*M* = .484, *SD* = .020) with WDNI (*M* = .553, *SD* = .020), and IDNI (*M* = .496, *SD* = .019) with WDNI (*M* = .553, *SD* = .020). The only reliable difference for LTE Segment2 is between the IDI (*M* = .349, *SD* = .016) and WDNI (*M* = .392, *SD* = .017) conditions.

A last test was run to examine correlations between Emotiv outputs and subjects' self-reported levels of engagement and mood. Correlations were only found in Segment1 of the IDI condition where self-reported levels of engagement correlated with STE (r = .233, p<.05) and LTE (r = .244, p<.05). No other correlations were identified among variables.

DISCUSSION

Analysis was conducted to assess the accuracy and stability of the associated Affectiv Suite outputs in a learning context and to compare them to self-reported measures. Before breaking down analysis results, it is important to address the limitations associated with the Emotiv in considering its applicability as a low-cost solution for informing student models. First, there is no clear definition of what the outputs are truly reporting. A basis for this study is to determine if their output values reflect the detection state they are defined within. Next, there is no indication of how noise in data is filtered out. When the device determines there is too much noise to calculate a detection state, the Affectiv Suite outputs all values of 1 until noise in data reduces. This spike in state values can have a significant impact on calculated means. These limitations must be considered when interpreting the justification of results.

Results from the study strongly support the hypothesis that the Emotiv can reliably differentiate brain activity between rest and active states. A visual representation of the metrics' associated trends (see Figure 3 on next page) shows consistency in metric outputs across all conditions and within all time segments. Interestingly, going from a rest state into task execution produces instant increases in engagement levels and reductions in excitement levels. Once scenario interaction begins, engagement stabilizes and holds over time while both excitement metrics significantly decrease between Time Segment1 and Time Segment2. This inverse relationship is supported by previous research investigating stress and control of performance (Matthews, Davies, Westerman, & Stammers, 2000). Through modes of compensatory control, an individual processing information compensates for any threats to

performance through active control and effort (Hockey, 1986). This finding supports Emotiv's ability to track engagement and excitement trends as they relate to learning events.

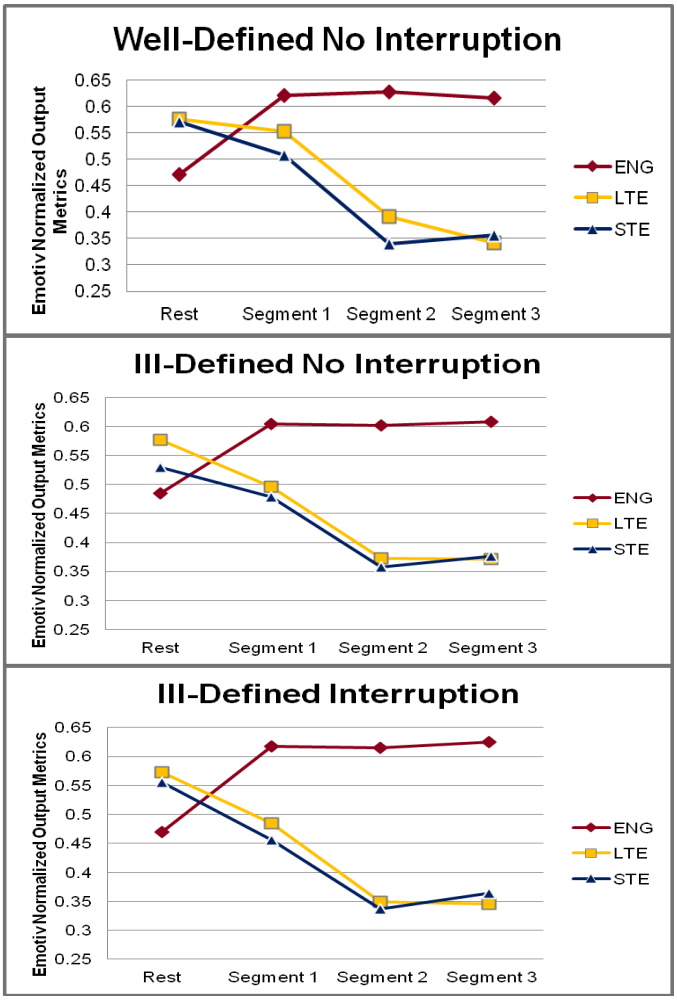

Figure 3. Graph of Means across All Time Windows for Each Scenario Condition

To further test the sensitivity in Emotiv metrics, repeated-measure ANOVAs were run to view the effect the independent variables (interruption and clarity of task) had on output metrics. It was hypothesized that the Affectiv Suite outputs would be significantly affected by the presence of an interruption in task flow. Results did not support this claim. It is the authors' opinion that an interruption in the pattern of conversation did not have an effect due to the static interactive environment and the nature of the task.

Additional analysis was conducted to observe the trend in metrics between conditions as a participant progressed through scenario. This is to see the effect time on task has on Emotiv outputs. Results show significant differences in the two excitement metrics when comparing well-defined against ill-defined conditions, where output values declined considerably faster in IDNI and IDI. This is supported by ill-defined tasks requiring more compensatory control of active attention and effort due to lack of clarity in task execution.

Overall, this study supports the use of the Emotiv as a low-cost solution to modeling cognitive state for desktop training applications. Additional research is required to assess the effect varying methods of task intervention have on cognitive engagement across multiple computer-based platforms, and to further test Emotiv's ability for detecting shifts specific to task engagement. This research can inform adaptive strategies to execute when cognitive function negatively impacts learning.

REFERENCES

American Electroencephalographic Society Guidelines for Standard Electrode Position Nomenclature. (1991). *Journal of Clinical Neurophysiology*, 8, 200-202.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., . . . 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), B231-B244.

Bradley, M. M., & Lang, P. J. (1994). Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1), 49-59.

Coyne, J. T., Sibley, C., Cole, A., Gibson, G., Baldwin, C. L., Roberts, D., & Barrow, J. (2010). Adaptive Training in an Unmanned Aerial Vehicle: Examination of Several Candidate Real-Time Metrics. In W.Karwowski & G. Salvendy (Eds.), *Applied Human Factors and Ergonomics*. Boca Raton, Fla: Taylor & Francis.

Craig, S. D., Graesser, A. C., Sullins, J., & Gholson, B. (2004). Affect and Learning: An Exploratory Look into the Role of Affect in Learning with Autotutor. *Journal of Educational Media*, 29(3), 241-250.

Emotiv Systems, I. *Emotiv Software Development Kit: User Manual for Release 1.0.0.3*. San Francisco, CA.

Fabiani, M., Gratton, G., & Coles, M. G. (2000). Event-Related Brain Potentials. In J. T. Cacioppo, L. G. Tassinary & G. G. Berntson (Eds.), *Handbook of Psychophysiology* (pp. 53-84). Cambridge, MA: Cambridge University Press,.

Hockey, G. R. J. (1986). A State Control Theory of Adaptation to Stress and Individual Differences in Stress Management. In G. R. J. Hockey, A. W. K. Gaillard, & M. G. H. Coles (Eds.), *Energetics and Human Information Processing*. Dordrecht: Martinus Nijhoff.

Kim, J., Hill, R. W., Durlach, P. J., Lane, H. C., Forbell, E., Core, M., . . . Hart, J. (2009). Bilat: A Game-Based Environment for Practicing Negotiation in a Cultural Context. *International Journal of Artificial Intelligence in Education*, 19, 289-308.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent Tutoring Goes to School in the Big City. *International Journal of Artificial Intelligence in Education*, 8, 30-43.

Lombard, M., Ditton, T. B., & Weinstein, L. (2009). Measuring (Tele)Presence: The Temple Presence Inventory. In *Proceedings of the Twelfth International Workshop on Presence*. Los Angeles, California, USA.

Matthews, G., Davies, D. R., Westerman, S. J., & Stammers, R. B. (2000). Stress, Arousal, and Performance. In *Human Performance: Cognition, Stress, and Individual Differences*. Philadelphia, PA: Psychology Press.

Norman, D. A. (1981). Twelve Issues for Cognitive Science. In *Perspectives on Cognitive Science* (pp. 265-295). Hillsdale, NJ: Erlbaum.

Shute, V. J. (2007). *Focus on Formative Feedback*. Princeton, NJ: Educational Testing Service.

Sottilare, R., Goldberg, S., & Durlach, P. J. (2011). Research Gaps for Adaptive and Predictive Computer-Based Tutoring Systems. In *Proceedings of the International Defense and Homeland Security Simulation Workshop (DHSS)*. Rome, Italy.

Tang, A., Biocca, F., & Lim, L. (2004). Comparing Differences in Presence During Social Interaction in Augmented Reality Versus Virtual Reality Environments: An Exploratory Study. In *Proceedings of the 7th International Workshop on Presence*. Valencia, Spain.

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-Aware Tutors: Recognising and Responding to Student Affect. *International Journal of Learning Technology*, 4(3/4), 129-164.

Woolf, B. P. (2009). *Building Intelligent Interactive Tutors: Student-Centered Strategies for Revolutionizing E-Learning*. Burlington, MA: Morgan Kaufmann.