Taylor & Francis
Taylor & Francis Group

Theoretical Issues in Ergonomics Science

ISSN: 1463-922X (Print) 1464-536X (Online) Journal homepage: http://www.tandfonline.com/loi/ttie20

Modelling a learner's affective state in real time to improve intelligent tutoring effectiveness

Keith W. Brawner & Avelino J. Gonzalez

To cite this article: Keith W. Brawner & Avelino J. Gonzalez (2015): Modelling a learner's affective state in real time to improve intelligent tutoring effectiveness, Theoretical Issues in Ergonomics Science, DOI: 10.1080/1463922X.2015.1111463

To link to this article: http://dx.doi.org/10.1080/1463922X.2015.1111463

THEORETICAL ISSUES IN ERGONOMICS SCIENCE, 2015
http://dx.doi.org/10.1080/1463922X.2015.1111463

Taylor & Francis
Taylor & Francis Group

Modelling a learner's affective state in real time to improve intelligent tutoring effectiveness

Keith W. Brawner[a] and Avelino J. Gonzalez[b]

[a]Army Research Laboratory, Orlando, FL, USA; [b]University of Central Florida, Orlando, FL, USA

ABSTRACT

This paper introduces, describes, and evaluates real-time models of affective states of individual learners interacting with Intelligent Tutoring Systems. Computer-based instructors, like human instructors, should use affective information for adapting instruction. This requires an accurate representation of individual learner state during tutoring; however, models described in the literature are generalised and constructed offline. Such total population models have faced validation difficulty with individuals, while individualised models have had difficulties with offline creation and online use. The simultaneous creation and utilisation of an individualised model from sensor-based physiological measurements presents an attractive alternative. We present and evaluate approaches for building affective models during the tutoring session which address the difficulties present in real-time data streams. Additionally, this work examines the impact of occasional direct user query on model quality. The results indicate that individualised real-time model construction is comparable to offline equivalents, yet can be successfully applied in tutoring settings.

ARTICLE HISTORY
Received 14 August 2015
Accepted 19 October 2015

KEYWORDS
Affective learner state; real-time systems; computer-managed Instruction; artificial intelligence; intelligent tutoring systems

Relevance to human factors/ ergonomics theory

The Theoretical Issues in Ergonomics Science journal serves to advance the science and philosophy of human factors and ergonomics through providing a vehicle for dissemination of research in the scientific foundations of human-centered and human-compatible systems. The research in the attached paper covers these issues directly in a number of manners. Firstly, it reviews the scientific foundation of human-centered and human-emulating tutoring systems. Secondly, it provides a review of the scientific underpinning of the educational systems aspects of affective computing. Finally, it combines these reviews in the research and design of potential solutions for enhancing future adaptive learning systems. These solutions attempt to optimize the system for performance while addressing the unique problems and opportunities that a human presence implies. The approaches and solutions presented in this paper are especially relevant to this special issue as they are the type personalized, and theoretically grounded, enabling technologies requested in the call.

1. Introduction and motivation

Tutoring by an expert human tutor is extraordinarily effective. Studies have found that tutored learners outperform their traditional classroom equivalents by between one and

CONTACT Keith W. Brawner keith.w.brawner.civ@mail.mil

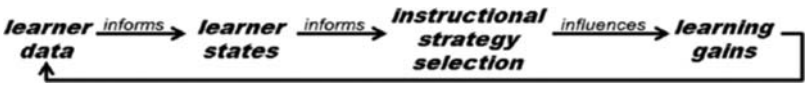

Figure 1. Learning effect chain (Sottilare 2012).

two letter grades of improvement (Bloom 1984; VanLehn 2011). The increase in learning resulting from these tutoring interactions has inspired a field of study in Intelligent Tutoring Systems (ITS) to emulate these results with machine-based tutors.

The hypothesis in the ITS field states that individualised tutoring can be provided inexpensively via computer and can be as effective in producing learning gains as one-on-one human tutoring (Verdú et al. 2008). However, this has yet to be shown unequivocally in the literature (Woolf 2009; Koedinger et al. 1997). The latest thrusts in ITS research deal with systems that are sensitive to the affective and cognitive needs of the learner in order to automatically implement an appropriate instructional strategy for that person at that moment in time during the tutoring session (Woolf 2009). This represents computer instruction in a way similar to how human tutors instruct – with attention to affect and cognitive state management (Kim and Baylor 2006; Lepper and Hodell 1989; Woolf 2009). Theory indicates that learner data inform learner states which inform instructional strategy selection which influences learning gains (Sottilare et al. 2012), shown in Figure 1.

Within the context of this work, these learner data refer to physiological data gathered via sensors, but in principle refers to any information about learners. The collection and interpretation of these data, especially their use for building models of learner states, is the main interest of our work and the subject of this paper. As ITS research moves towards highly adaptable and individualised tutoring, the need to automatically assess the cognitive and affective states of the individual learner for instructional adjustment has been well documented (Army 2011; Woolf 2010). Extensive work has been performed to recognise the emotional state of a learner by incorporating sensors to monitor both behavioural and physiological markers and is discussed in the coming sections (D'Mello, Taylor, and Graesser 2007; Berka et al. 2007; McQuiggan, Lee, and Lester 2007; Sidney et al. 2005; Beringer and Hüllermeier 2006; Baker et al. 2012; Chaouachi and Frasson 2010; Cooper et al. 2010; Conati and Maclaren 2004).

However, the prediction and classification of these affective and cognitive states from a stream of physiological and behavioural data has proven difficult. The authors believes that there are several reasons for this difficulty: (1) the models used generalised data obtained from a large sample of human subjects, which lacks applicability to individuals; (2) poor classification accuracy for single-user individualised models because of changes in individuals over time; and (3) difficulties in collection of appropriate datasets. Each of these difficulties, along with the research findings from various projects, is discussed individually in the coming sections. As a byproduct of these difficulties, affective data mining for educational purposes, especially from physiological data, has been slow to gain momentum, resulting in limited validation of differing modelling approaches. Furthermore, the fitting of generalised models (e.g. of a population) into use for individual learners is one possible reason for the general difficulty experienced in transitioning such models from research into actual use, and that individualised modelling and direct user self-report query may be a good method for enhancing effectiveness and transition.

Our research questions, therefore, relate to (1) whether individualised affective models, created and utilised concurrently in real time during a tutoring session, perform better than their generalised and offline counterparts, and (2) whether the model building process is positively supplemented by self-reported ground truth data in real time while the learner is being tutored. Simply said: can real-time models be created and can they be enhanced via occasional user query?

Before presenting our current work, however, we review previous attempts to model affective states of learners in order to frame the problem of real-time state modelling. While the literature contains many reports of *generalised* affective state models, *individualised* models of affective states have been only rarely mentioned. In Section 1, we present a review of both generalised and individualised affective modelling research for training and education while attempting to answer the question of suitability of individualised models. In Section 2, we discuss the desirable features of a dataset for building useful affective models. Section 3 contains a description of the machine learning techniques we employed in building such real-time models and how we modified these techniques to support active learning. Sections 4 and 5, respectively, describe the experiments and their results, and the conclusions to be drawn from them.

The following sections describe the state-of-the-art in building models of learner affect and cognition for use in ITS systems. Note that our work only pertains to affective models and not models of cognitive states. However, because both are important in tailoring instruction in ITS, and because they are frequently modelled using the same techniques for the same purposes, our review covers both types of models. This joint work between the fields of physiology, human factors, instruction, and computer science represents steps towards the application of the various technologies possible and employed.

1.1. Related research – generalised modelling

The first step to tailoring a pedagogic strategy in an ITS that can respond to affect is, naturally, to detect and identify the affective state of the individual learner in real time. While self-reported data can be used for this task, these operations can ideally be performed in a manner that does not interrupt the student from the learning task, thus consuming both time-on-task and student cognitive resources. Affective state identification has been performed from a variety of interaction and physiological signals; however, for practical reasons, these signals must have validity insofar as the affective state, reliability, appropriate time resolution, and be minimally intrusive. Previous work (Calvo and D'Mello 2010) has found physiology to be relevant to affect detection, with relevant variables involving heart rate, skin conductance, motor response, facial coding, skin temperature, or other measures.

The idea of using an emotionally responsive ITS to improve learning is not new. After all, this is what human tutors regularly do. The focus in the early days of affective educational modelling was on building general models of learner state that would predict the state of a population of learners over an arbitrary period of time. This model construction is followed by a period of validation, where the ability of the models to accurately predict learner state is tested with other populations (or the same population at different a time or under different conditions). Conati (2002) reported the first attempt of this process in 2002 when probabilistic models of emotion through the interaction with learning systems were used in a game called 'Prime Climb', an educational number factoring game with

hinting based on probabilistic knowledge representations. Data were measured via bodily sensors to construct a model of emotions, theorised to be hidden behind interaction data (i.e. students in different states interact differently). However, the approach first experienced a problem that was later found to consistently plague generalised affective models: difficulty in validation process. In the case of the Prime Climb system, Conati and Maclaren attempted to validate an *offline*, *generalised* approach based upon Dynamic Bayesian Networks (DBN) (2004) through measurements of predictive accuracy, which experienced high degree of error, likely as a result of the aforementioned differences among individual learners.

Craig et al. (2004) investigated the influence of affect during tutored interactions using a tutoring system called AutoTutor. AutoTutor can be used to instruct a variety of subjects through assessment of dialogues and speech acts based on similarity metrics to preconfigured scripts; it is commonly used to teach principles of physics. The studies of affect in AutoTutor were focused on the detection of states to correlate with learning gains in a research setting, rather than for their use in a production system. Sensor measurements for affect detection included body posture, keyboard pressure, and mouse pressure (Sidney et al. 2005). The collection of sensors used was comparatively extensive in relation to other such studies. However, similarly to Conati's findings, this study experienced difficulty in validation of *generalised* models created *offline* (Graesser et al. 2007).

Mott and Lester (2006) investigated the inclusion of sensors for affect detection in the Crystal Island ITS. Crystal Island is a tutoring system for teaching middle school biology concepts, such as disease causes, via student interactions and explorations of a 3D game world with a variety of agents. The classification approach of this system made use of measures of temporal interactions, location features, intentional features, physiological response from blood volume pulse and galvanic skin response (GSR). These measurements were collected and classified using various machine learning algorithms (McQuiggan, Lee, and Lester 2007), including Naïve Bayes, decision trees, support vector machines (SVMs), and *n*-grams, all of which showed a predictive accuracy superior to baseline conditions. The authors applied many different *offline* machine learning algorithms in an effort to create *generalised* models. These offline generalised models appear useful, as they have high classification accuracy, but attempts by these authors to validate their models failed, thereby stressing a continuing need for validation of their models. These validation attempts are discussed in greater detail within the next section.

1.2. Related research – generalised validation

In summary of the previous section, studies conducted with generalised modelling techniques show that it is *possible* to build models to recognise affective states in human users, but that the models do not meet validation thresholds at a generalised level. A few researchers have pursued efforts directly intended to create and then validate affective models in different populations, but like the projects described in the previous section, these efforts also failed to successfully develop generalisable, offline models. This section discusses the attempts of these authors to validate their results in greater detail, as there is much to be learned from the few existing validation studies.

Sabourin, Mott, and Lester (2011) continued to study affect in the context of the Crystal Island ITS through the investigation of generalised affective models from system

interactions. Their 2011 study is one of only two published research articles with validation results, rather than merely mentioning that validation was difficult, as did all others before it. Sabourin et al. reported data from 260 learners from two schools. Their study included injection of experimenter knowledge of student tasks into the models in an attempt to eliminate statistical options and aid in algorithmic (Dynamic Bayesian Network) performance. The use of experimenter knowledge during model creation for educational purposes is undesirable, as it makes it highly unlikely that the model will transfer to another domain of instruction. In this study, the models created from data at one school dramatically underperformed baseline measurements during validation on a second school. The authors conclude with the statement that although 'models were evaluated in a subject-independent manner, they were not successfully able to extend to a future population. This finding is particularly interesting given the strong similarities between the two populations' (Sabourin, Mott, and Lester 2011).

A second study that attempted validation was published by Cooper et al. (2010). They conducted the study in a similar time frame to the other previous studies (McQuiggan, Lee, and Lester 2007; Sabourin, Mott, and Lester 2011), and report 80-90% accurate classification of affective state via Bayesian networks that gathered data from students using webcam-provided Facial Action Coding System (FACS) information, posture sensing devices, skin conductance, and a pressure sensitive mouse. This research eventually converged on a set of sensors used for several later studies. The dataset informing these models was collected and labelled on 100 students in the Fall semester and used in an unlabelled setting in the Spring semester with 500 students. This study indicated that there was no validated accuracy above baseline when attempting to transfer these highly accurate *generalised* models, created *offline*, into practical application. In short, despite similarities among groups, group-created models have not seen prediction accuracies above chance when applied to a different population.

In both of these notable studies (Sabourin, Mott, and Lester 2011; Cooper et al. 2010), their authors were able to put into practice systems that appeared functional. However, these generalised emotional models built with several offline machine learning methods barely performed better than baseline. Even worse, they were shown to not transfer to real-world conditions.

This evidence points to a significant gap in the research. While there have been many studies that *correlate* physiological data to various experiences among groups, there have been few studies that *use* models based upon physiological data to make real-time decisions. This leads to the hypothesis that individual differences between subjects may be the root cause of the failure of these generalised models to transfer. The problem of individual differences between human subjects forces a researcher to consider each person individually (e.g. using an *individualised*, rather than a *group*, approach). These types of approaches are discussed next. They relate to the first research question stated in Section 1.

1.3. Related research – individualised approaches

Individualised approaches to affective data analysis are rare, but not completely absent from the literature. Furthermore, certain authors of generalised modelling publications have pointed to individualisation as a possible solution (Calvo and D'Mello 2010). Certain types of signals, such as electroencephalography (EEG), naturally lend themselves to

individualised approaches. As human brains are highly individual, and consequently, brain models produced from EEG signal data are also highly individualised. EEG studies have hinged upon the development of highly individualistic models, and therefore also provide an example of a field where individualistic analysis is commonplace.

Another viable approach involves sensor suites. For example, Blanchard, Chalfoun, and Frasson (2007) hypothesised that a combination of sensors could identify a user's emotional state without bias, and successfully account for individual differences within the data in the context of training and education. They used a combination of sensors for the measurement of physiological state, including skin temperature, respiration, heart rate, blood volume pressure, GSR, surface electromyography (EMG), and EEG. Their models, while individualised, were built offline. The authors were critical of the use of post-hoc analysis, and highlighted the need for a real time or 'predictive model' approach that is able to quickly classify learner state, given a set of sensor inputs, for use in real-time pedagogical adaption, e.g. an online model construction approach capable of immediate use during the training session.

Blanchard et al., however, cited difficulties prevalent with large individual differences in physiological data. As an example of individual variations experienced in response to the same situation, they include a figure that shows a 10% change ratio in absolute GSR measure (range 4.9−5.3) for one individual during a training session, while another individual experienced a 133% change ratio (range 5−13) using the same sensor and placement. Similar individual variations in sensor ranges were found for EEG, skin temperature, and other measures, even after filtering, highlighting the problem of individual variations: the variance from one individual to another, using the same sensors, can be large. The authors mention that the construction of individualised models is 'not only possible but highly recommended', and suggest firmer techniques for individualised base-lining (Chaouachi and Frasson 2010).

In AlZoubi, Koprinska, and Calvo's (2008) early research into EEG models, participants were taught to play Pong by thinking of moving their left and right arms while connected to an EEG measurement system. After this, a model of left and right arm movement was constructed for each participant. The participant then had to think of left and right arm movement in order to control a virtual cursor. The findings related to this work were that models were highly individualised and that the best offline classification system was *never* the best performing (e.g. validation) classification system; the models with highest classification accuracy in an offline setting were never the models with highest classification when put into use. Furthermore, they found that *individualised and offline* classification models experienced sharp decrease in accuracy when used for EEG classification during live tasks (AlZoubi, Koprinska, and Calvo 2008). These findings are consistent with the findings presented for *generalised* models described earlier in this paper, which experience the same sharp decrease during the validation phase of the research.

Further work in this area by AlZoubi, Calvo, and Stevens's (2009) indicates that affective state classification is possible from the EEG sensor array, but cite significant difficulties arising from user fatigue, electrode drift, changes in electrode impedance, and user cognitive state modulation. They argue that the problem inherent in these physiological signals is their non-linear nature and that the failure of other models is because of the underlying linear assumptions. Each of these models, when created offline for online use,

assumes that the user will be in similar state, with similar baselines, at similar fatigue levels, with millimetre accuracy in placing electrodes, etc. They indicate that the models are poorly fit for use when assuming that the underlying concept is stationary, when in fact it is drifting across the sampling space (Hulten, Spencer, and Domingos 2001); models should be adaptive and continuously adjusting for the reasons enumerated above as well as others. As such, they hypothesise that nonlinear algorithms could successfully deal with the dynamic nature of the signal. AlZoubi, Calvo, and Stevens empirically show this success through an injection of real-time adaptive algorithmic techniques, such as *windowed* Bayes Networks, which produced 40% less overall error (2009).

With their adaptive approach, AlZoubi et al. addressed the problem of day-to-day individual differences in multichannel physiology (2011). They concluded that based on a laboratory study with induced emotions that it would be possible for such an approach to be implemented in the field. While their approaches using *offline* adaptive algorithms are not particularly suitable to the problems of real-time classification, they present a picture of the problems faced in their current work. This helps to clarify our original research questions involving individualised models to a number of sub-research questions:

(1) Can individualised affective models, created with online and adaptive machine learning approaches in real time, perform comparably with their generalised counterparts? Can they be validated in a person-independent context?
(2) What are alternative methods for online and adaptive classifiers?
(3) Is the model building process positively supplemented by self-reported ground truth data in real time while the learner is being tutored?
(4) How can these algorithms be applied to sparsely labelled data?

1.4. Online real-time models

The use of learner affective models is still among the most promising technologies for the tailoring of individual training, as indicated by Woolf (2010). The previous sections have shown a need for *validated* models that can predict/classify the affective state of the learner in real time, during the tutoring session. The creation of a validated affective model that is both adaptive and individualised at runtime presents an opportunity for it to be transferred into operational ITSs that can use this model to better inform instruction. The first step towards the use of the prediction/classification of affective models is to evaluate the effectiveness of the approach.

The clearest and most natural way to evaluate the effectiveness of any model is by assessing its accuracy. However, there are problems with evaluating the type of models of interest here merely based upon their accuracy. Not the least of these problems is that no study exists that links affective model accuracy with learning gains in a tutoring context. Of course, this is because current affective models have not been accurate, or if accurate, these models have not successfully transferred. This disconnect further stresses that models should be built for their *use* rather than for their predictive accuracy, as the end goal of an ITS is to *improve instruction*, rather than to provide accurate *student assessment*, although they are theoretically related.

Before running various experiments, it was thought that the real-time model construction approach may potentially sacrifice overall accuracy in exchange for the real-time availability of predicted values. This would be acceptable in light of the aims of our

work – to transfer the models to an operational environment. In short, a model able to adequately inform an instructional decision *while the student is in need* has more value than a more accurate model at a less relevant time, such as after a student has ended a training session.

Therefore, the hypothesis of our research is that more useful affective learner-specific models can be constructed in real time during the training session. We further hypothesise that these individualised models of affect, created in real time, can achieve accuracy on a par with, although possibly slightly diminished, individualised offline models created for the same learner. Our work directly addresses the challenge put forth by Calvo and D'Mello (2010) with regard to affective categorisation and affect detection system evaluation.

2. Data acquisition

To build these models, of course, one needs data. In this section, we discuss the data used to build the individualised models in real time. We begin by describing an ideal dataset, as useful datasets for this sort of work are rare. Since the conduct of this experiment, the authors have used the below checklist as a guide to the creation of future datasets, intending to carry out further research in this area. The description below of an ideal dataset is intended to aid future researchers.

2.1. Ideal dataset description

Access to data for building affective learner models can be difficult, as the availability of a context-appropriate datasets is limited. The creation and sharing of such datasets is an area where the field has been lacking, and research into this area could benefit if an open standard for such data were to be adopted. An ideal dataset for creation of an educationally based model of affective state of a learner includes several features, which are identified as follows:

(1) Relates to affective states relevant to learning.
(2) Ability to be used in applications other than the system of creation (i.e. the same data with the same sensors can be used in another educational setting).
(3) Obtained from a relevant population (i.e. learners learning).
(4) Obtained using cost-appropriate sensors.
(5) Contains labelled data (otherwise, it is impossible to validate).
(6) Has been used in previously established models to be used for comparison (i.e. allows for ablative studies).
(7) Was collected in a relevant setting.

Many studies meet one or more of these criteria – collection using cheap sensors in a classroom meets the majority of the criteria. However, many affective datasets do not address all of them, with #1, #2, and #6 being the most commonly omitted criteria. An example of a dataset that does not meet criterion #1 is the Pose, Illumination, and Expression database (Gross et al. 2010), which shows actor expressions. Criterion #6 is frequently left unmet through the research need to assess sensors, interventions, or other items, rather than the ability of various methods to model or predict.

With respect to criterion #2, there have been several studies that collected emotional data that were only transferable to similar learning environments. One

example is Baker et al.'s (2012) dataset, which draws emotional inference based on the actions that the student takes within a learning environment. Such data cannot be transferred to another learning environment that features different actions. Another example is data about 'gaming the system' models. These models predict whether the student is meticulously studying based on his/her interaction with system-dependent screen elements (Baker et al. 2004). This type of model is referred to as an interaction-based model, which can be contrasted with models based upon collection of sensor data. Sensor-based models have transferability, as a sensor can supplement an existing system, while interaction-based models are dependent on the system of interaction. Sensor-based models are of interest to our research, as we believe it will address the many needs of ITSs.

2.2. Dataset acquisition

The dataset determined to most closely match the above collection criteria, and the one used in our study, was part of an experiment to evaluate low-cost sensors. While these data were collected by the first author, they were not collected for the purposes of this study. Information about the data collection and public availability is found in Brawner (2014), but a summary is discussed here. In this data collection process, college-aged military learners experienced a breadth of learning-relevant emotions while watching videos or playing video games, while they were measured by a suite of sensors. Cognitive states, such as distraction, were labelled with a high-cost sensor. Affective states, such as frustration, were labelled with a self-reporting tool, which has the 'high cost' of frequent user query, using time that could be better spent learning. Models developed in our research were designed to replace the high-cost EEG sensors' measures as well as the time-consuming affective self-reporting.

The baseline measure of the affective portion of this dataset is the EmoProTM validated electronic emotional profiling tool (Champney and Stanney 2007), which identifies the affective state after a brief period of questioning following an emotional episode. Briefly, the research question addressed in the original study that produced this dataset was 'Can you replicate the measures of validated, high-cost, obtrusive sensors with yet-to-be-validated, low-cost, unobtrusive ones?', where the low-cost sensors are defined by the list below, and described in greater depth in other works (Brawner 2014; Carroll et al. 2011; Kokini et al. 2012; Brawner 2013). These devices and the measurements used are briefly described in Table 1.

(1) Custom eye-tracker
(2) Zephyr Heart Rate Monitor
(3) Phidget-based Chair Pressure Sensor
(4) Vernier Motion Detector
(5) NeuroSky MindSet EEG

A power analysis conducted for this study determined that 18 participants were necessary to determine which of the sensors could reliably gather affective and cognitive state information from the participants. Twenty-seven (27) datasets were collected; however, only 19 of these provided usable emotional labels with all their sensors providing stable information across all of the events. Each sensor used for this study was selected because of its low cost, which unfortunately also correlated with low reliability. The eight discarded sets of data were rejected because one or more of the sensor data streams became

Table 1. Summary of sensor measurement (see Brawner 2014 for details on these sensors).

Sensor	Measures	Variables
EmoPro (ground truth)	Anger Anxiety/Fear Boredom	Self-reported Boolean values
NeuroSky EEG	Alpha1, Alpha2, Gamma1, Gamma2, Delta, Beta1, Beta2, Theta, Attention, Meditation	Various measures of brain activity in the specific frequency band, from the two sensor locations. Two derived measures, based on NeuroSky software algorithms
Zephyr HRM	Heart rate	Reported as it changed
Vernier Motion Detector	Motion	Reported distance from the laptop computer
Chair Press sensor	Chair 1−8	Report the pressure on four locations on each the back and the seat.
Custom Eye Tracker	Left Eye Pupil Diameter	Pupil Diameter
Difference-based features (software creation, simple difference between last and current value)	Alpha1 Diff; Alpha2 Diff; Gamma1Diff Gamma2 Diff; Delta Diff; Beta1 Diff; Beta2 Diff; Theta Diff; Attention Diff; Meditation Diff; Heart Rate Diff; Motion Diff	Calculated features using the simple formula, Diff = current-previous formula.

unavailable due to malfunctions. This rendered it impossible to evaluate which of the sensors contributed to a generalised model of affect during offline analysis for that particular individual, or provide a point of comparison during online analysis. In a real-world setting, a system should be able to respond to the lack of availability of one or more input data sources. Nevertheless, having 19 individual models satisfied the requirement of 18 sets of individual data.

The population of interest was United States Military Academy (USMA) cadets, with nine to 44 months of experience at West Point. This is roughly equivalent to a population of modern college students. The majority of the members of the population were plebes (first year learners) enrolled in the Behavioral Sciences and Leadership (BS&L) Department's General Psychology (PL100) course.

Participants were asked to undertake a visual vigilance task, watch video clips from the movies Halloween and My Bodyguard, and play several scenarios within the Army's Virtual Battlespace 2 (VBS2) video game. The video segment from Halloween has been previously validated to induce Fear/Anxiety, while the video segment from My Bodyguard has previously been validated to induce Anger/Frustration. The VBS2 scenarios contained limited visual perception (validated to produce fear, anger, workload), large numbers of enemies (validated to produce fear, anger, workload, and engagement), annoying sounds (validated to produce anger, workload, and distraction), and equipment malfunction (validated to produce anger, fear, workload, and distraction). More information about the conduct of the study and initial analysis may be found in its original publication (Kokini et al. 2012), while other sources better describe the initial affect elicitation validation of VBS2 scenarios (Jones et al. 2012) and movie clips (Hewig et al. 2005).

This dataset is not ideal, but is as close to the ideal as can be currently found. Of the seven requirements discussed earlier, it satisfies six of them. It has learning-relevant states, from a learning population, with affordable sensors, labelled data, previously established benchmarks, and should be able to generalise emotion detection beyond its construction system. The only requirement not addressed is #7 − collected in a relevant setting, as it does not contain data during interactions with learning events. Furthermore, the data used in this study was obtained in a laboratory, with a small sample size, using artificial

Table 2. Summary of reported state instances (dataset provided as supplemental material publicly available with documentation; Brawner 2014).

Self-report	No. of times reported
Anger / no anger	60 / 249
Boredom / no boredom	37 / 272
Fear / no fear	39 / 270

mood induction rather than in the field. Nevertheless, it is the best set of data that was available to us, and notwithstanding the above, it is close to being ideal.

After each of these events, the participant was affectively measured with the use of the EmoPro self-report tool to periodically provide 'ground truth' of affective state. The data from the experiences were labelled to be of the self-report class (e.g. anger, boredom, frustration). The EmoPro labels represented several minutes of real time prior to a single label and correspond to a large number of data points. Events were kept short to increase the resolution of the EmoPro data, totalling 309 labelling instances for each of the 19 users, with states summarised in Table 2. While cognitive states were also labelled with the ABM EEG headset, the work reported in this paper focuses on the affective states of the learners, which did not use the EEG data.

2.3. Offline model benchmarks for comparison

Carroll et al. (2011) previously analysed the results of this experiment offline to determine how well the combined sensors were able to detect the labelled affective state of the learner. With regard to our work, these analyses serve as benchmarks for comparisons (e.g. ideal dataset criterion #6). These models built by others represent the best effort by other researchers to build models of affective states via offline approaches.

The Logistic Model Regression method was integrated in the Logistic Model Trees technique that was selected as the method to use to create the models. 10-fold cross-validation was used to prevent model over-fitting. The sample was then analysed with the receiver-operating characteristic (ROC) function benchmark (Hanley 1989), which plots the proportion of correctly classified observations from the positive class (true positive rate) against the incorrectly classified observations (false positive rate). The *Area Under the Curve* (AUC) of this function was calculated as a direct measure of the performance of the resulting model. The AUC ROC is designed to compensate for the misleading figures of 'percentage accuracy' in unbalanced data. The AUC ROC measurement allows an algorithm with lower overall errors, either false positive or false negative, to score well (Hanley and McNeil 1983), as all the categories of possible classification are weighted equally. In general, AUC values of greater than 0.8 are considered to represent good performance, while classifiers lower than 0.6 are considered poor; scores between 0.6 and 0.8 are considered acceptable performance. It is mathematically defined below, with the TruePositiveRate being a sample of the model's accuracy, FalsePositiveRate being a sample of the model's inaccuracy, and dModel being an incremental sampling adjustment:

$$\int_{\infty}^{-\infty} \text{TruePositiveRate}(\text{Model}) * \text{FalsePositiveRate}(\text{Model}), \text{dModel} \ .$$

Table 3. Previous benchmark results using logistic model trees (Kokini et al. 2012).

	Affective measure		
	Anger	Anxiety/fear	Boredom
AUC ROC value	<0.6	0.83	0.79

These other models were built offline with *no limits on the time* available to build these models. Furthermore, each of these models is constructed with *all* of the data available, and with *all* the true class labels. With all data, all labels, no time limit, and well-reasoned research approaches, these generalised models can be said to represent the 'gold standard' against which to later compare our online, real-time models that did not enjoy these advantages because of their real time, on-line constraints. As mentioned in Section 1, these offline and generalised models are not expected to transition to the field because of the generalised vs. individual nature of the models. Nevertheless, they do represent the highest accuracy in classification that could be obtained. We would not expect that our online model with limited data, limited label availability, and significantly constrained time to create the model would result in superior accuracy to these benchmark models. This represents the engineering trade-off of accuracy for individualised and availability previously discussed. The results of these offline affective state models are previously reported by Kokini et al. (2012), are re-printed in Table 3, and serve as a benchmark for comparison in the creation of online models later in this work.

An interesting aspect of the above work of Kokini et al. is that only some of the features of the total data stream were used in their offline-created models. In short, the model for Anxiety/Fear used all of the sensors, while the model for Anger could not be created with predictive quality above baseline. The regression-based model of Boredom used only two of the available sensors: the NeuroSky EEG and the Zephyr Heart sensor, while ignoring the other sensor data feeds. Nevertheless, given that the offline modelling efforts achieved greatest model quality when ignoring some of the sensors, comparisons using both a Boredom subset (NeuroSky/Zephyr) and full set (all sensors) may still be viewed as equivalent.

3. Methods used for real-time classification

The research questions addressed by this work all relate to the suitability of real time, on-line classification of affective data streams. Additionally, we also sought to investigate potential advantages obtained from the reduction of total labels required. Given the topic of research on online real-time models, it is reasonable that only algorithms that can deal with the challenges of real-time computing are able to address our research needs. These challenges are divided into four main areas (Beringer and Hüllermeier 2006). In brief, these are (1) the data can be of potentially infinite length, (2) concept detection (i.e. identifying a group of data as a state), (3) concept drift (i.e. expansion of a representative group to encompass additional data), and (4) concept evolution (i.e. representing a single state in multiple/unrelated data organisations). Any method of real-time model construction requires addressing these challenges, which imposes serious design constraints, including but not limited to:

(6) The data stream is treated as infinite in length.
(7) The data stream cannot be stored.
(8) Data stream elements are not available for request.
(9) Presentation order is not controllable.
(10) There are strict time constraints, where all operations must be completed before arrival of the next data point.
(11) Classification is made at each time step.
(12) Decisions are based on encoded knowledge.
(13) New concepts, such as new affective states, must be identified quickly and tracked across a sampling space, based only on the encoded knowledge.

To answer the question of whether an online, data stream-based approach lends itself well to the problem of building real-time affective models, the literature was reviewed for real-time appropriate learning algorithms. The literature review addressed the state of the art for each real-time algorithm category. After this review, we converged upon a minimal but inclusive set to test, including: a clustering algorithm, a neural network-based approach, a graphical model, and an incrementally updating linear regression technique. One algorithm was selected as the state of the art in each of the categories of approaches, and each approach is sufficiently different from the others to warrant their specific inclusion in our study. We cover all fundamental approaches (although not necessarily specific adaptations thereof) covered in modern literature reviews (Jain 2008). Our work was undertaken with the goal of showing that the individualised real-time approach is valid as well as transferrable to practice. Each of the four algorithms is described below, along with, and the manner in which they address the above challenges and design constraints.

3.1. *A modified* k*-means clustering algorithm*

As Jain (2008) states: 'Organizing data into sensible groups is one of the most fundamental modes of understanding and learning'. Clusters are traditionally evaluated for fitness based on a distance metric. Clustering represents a standard approach for dealing with data of an unlabelled class and is the baseline method attempted as part of our work. A version of the *k*-means algorithm, modified for online fitness, was used and is described as follows:

For each new point, incrementally
Compare each point to all known centroids
If no cluster is within range of ⟨vigilance⟩
this point is a new centroid
Else, move the matched cluster ⟨delta⟩ in new point direction
Merge closest centroids based on ⟨vigilance⟩, if appropriate
Keep track of the number of points in these centroids, and the last point which modified the centroid, label if possible

3.2. *Adaptive resonance theory (ART) neural network algorithm*

ART is a type of neural network architecture that classifies objects based on the activation within a layered structure of recognition nodes. It was developed to classify data in a one-

pass learning process (Carpenter and Grossberg 1995). ART has a performance roughly equivalent to conventional feed forward neural networks, but with significantly reduced training time. In its most basic form, ART draws n-dimensional hypercubes around similar input patterns, where n is the dimension of the input data. Matched data are those that fall within the smallest hypercube or of the class of the closest available hypercube. Hypercubes are expanded to compensate for new data in accordance with parameter settings. The locations of the hypercubes are encoded as weight vectors. Although sometimes viewed as a disadvantage, the one-pass learning ability of ART systems makes them appropriate for real-time classification problems. This feature of ART makes it sensitive to the order of the input data. We anticipated that this would assist in the classification of affective computing signals, where the order of the input data is relevant to the affective signal, as shown in experiments with sensitivity (Carpenter and Grossberg 1987). The algorithm used in this work is described as follows:

For each new data point
Compute each neurons' weighted activation to it ($y\mathrm{i} = \Sigma w_{ij}{}^{*} x_i$)
Select the neuron with the highest activation
Test if this neuron in vigilance (x_i fuzzyAnd $w_x <$ vigilance)
If it is, update the weights:
$w_i = \mathrm{learningRate}^{*} x_i + (1 - \mathrm{learningRate})^{*} w_i$
Otherwise, create a new category with x_i weights

3.3. Growing neural gas (GNG) graphical algorithm

Growing neural gas (GNG) is a robustly converging alternative to the k-means approach of clustering that finds optimal representations based on feature vectors. These feature vectors construct a topographical map overlaying the data. This approach has its roots in self-organising maps (SOMs) and neural gas topologies. GNG is an incremental version of neural gas which is appropriate for data stream analysis (Holmstrom 2002), and was initially proposed by Fritzke (1995). Semi-supervised GNGs are a further outgrowth of these methods to make use of unlabelled data points for classification (Zaki and Yin 2008).

GNG is a relatively new technique for pattern recognition. It has seen increasing use in image recognition (García-Rodríguez, Flórez-Revuelta, and García-Chamizo 2007) and topology learning (Prudent and Ennaji 2005). Our previous research has revealed that it responds well to the injection of uniform noise information (Brawner and Gonzalez 2011). Fundamentally, the GNG algorithm creates an overlay to the data that detects edges in patterns and forms the areas interior to the edges into clusters. The boundary edges clusters serve to identify unique groups of data among the dimensions of the input space.

Beyer and Cimiano (2011) modified the initial GNG algorithm to remove its dependence on the Expectation Maximisation solution set, making it appropriate for real-time problems. They present Online, Semi-Supervised, Growing Neural Gasses (OSSGNG) as a topographical mapping algorithm synthesised from the various contributing fields. They examine several metrics for determination of the establishment of clusters, and find that the minimum distance metric has the best performance on problems of interest. We use the metric recommended by Beyer and Cimiano, but use the originally described

algorithm (Beyer and Cimiano 2011) without significant modification. The algorithm used in this work is described as follows:

Present a new point and find the two closest items (s_1 and s_2)
Increment the age of all edges coming from s_1
Compute the local error of s_1
error = squared distance from weight to input)
Move s_1 and its edge-connected nodes towards x_i in two ways:
Directly connected nodes: $\Delta w = e_b(x_i - w_{s1})$
Indirectly connected nodes: $\bigcirc w = e_n(x_i - w_{s1})$
If s_1 and s_2 are edge-connected, set the age of the edge to 0
Remove all edges older than the maximum age
If a node has no edge now, remove it
If it is time to present a new node:
Determine largest error node network from earlier calculated local errors
Determine the largest error point node in this network
Insert a node halfway between these two items, create edges, remove previous
Decrease all error by a factor, Alpha

3.4. Vowpal Wabbit (VW), linear regression

The previous methods discussed typically favour accuracy from among the various engineering tradeoffs. Vowpal Wabbit is a software package implementation developed by Langford, Li, and Strehl (2007) designed to be *fast* and use *as little data as possible*, with the assumption that labels are available. It makes extensive use of gradient descent and multiple passes over the data to train a variety of encoded weight vectors. The background assumption to our initial problem is that the data of interest are too voluminous to process efficiently, and that rapid training is critical. This approach was developed specifically for large-scale search operations. We included it to represent incremental and semi-supervised linear regression modelling approaches. It was used with modifications to support adaptive learning weights to compensate for limited training phase, and hinge loss function, as it is shown to increase classification accuracy in a constrained environment with binary classifications (Rosasco et al. 2004). The algorithm used in this work is described as follows:

Start with *for all i:* $w_i = 0$ *Within the loop:*
Get an example: $x \in (\infty, \infty)$
Make a prediction: $y = \Sigma i\ wixi$
Learn the Truth: $y \in [0,1]$ *with importance I*
Update the weight: $wi = wi + 2\eta(y\text{-}yi)I$
Repeat for specified number of passes or other criteria.

3.5. Learning approaches used – supervised, unsupervised and semi-supervised with active learning

There are several ways that the algorithms described above can be implemented. The machine learning community has traditionally segmented on the ideas of 'supervised learning' (with labels) or 'unsupervised learning' (without labels). However, a new field known as *semi-supervised*, or *transductive learning* (Zhu 2005), is beginning to emerge to

address some problems in each of the other fields, such as overtraining and supplementation of knowledge of clusters. Semi-supervised methods use information contained in the unlabelled data to (1) make inferences on the structure of the labelled data, and (2) reprioritise the classification of prior data points. Active learning methods use logic to select which points should have labels requested in order to accomplish either of these two goals.

Each of the methods used in our work is screened for its ability to deal with all the problems of real-time data classification, and the ability to handle the real-world issue of limited label availability. If an algorithm did not have an implementation for semi-supervised active learning, we created one for it. The most important feature of each algorithm selected is its ability to deal with the real-time data problems discussed above. We expected that some information about the user might be available during runtime, regardless of the level of supervision being used in model creation. The user may be asked directly about their state, but only occasionally, and this information can be used to help build a model.

Generally, the algorithmic implementation of the tests in this work relies upon the idea of a 'label request', where the algorithm is allowed to request (and be provided!) a finite number of labels in the course of model construction. Each algorithm must either maintain a list of likely hypothesis points that can disambiguate classification boundaries, or be able to generate one quickly.

The semi-supervised active adaptations are different for each algorithm because of the fundamental differences in how the data are represented within each construct. Generally, when given the choice to request a label, each algorithm attempts to label the largest body of unknown information. The clustering algorithm was modified to support semi-supervised active learning by responding to a label request from the evaluation framework by requesting the label closest to the centroid of the largest unlabelled cluster. ART was modified to support this requirement by disallowing known mixed-class classifications (when labels are plentiful) and responding to label requests with the largest currently unlabelled weight vector. The OSSGNG model was modified to request the largest currently unlabelled classification category and propagate labelling information across the category. VW supports semi-supervised active learning by default, in a mathematical representation of hypothesis likelihoods (Beygelzimer et al. 2010).

4. Experiments, results and conclusions

Before discussing results, it is useful to briefly discuss how data points of accuracy were generated and which measures of accuracy have been included. Generally, the manner in which performance values were generated follows the below algorithm:

For x from 10−100, in increments of 10
Feed x% of the data to the algorithm
For each class created by unlabeled class boundaries
Label this class the majority label of true set
Evaluate for AUC ROC accuracy through input of data for classification (next, previous, all seen)
Destroy model, loop

Three types of AUC ROC measures were taken for three slightly different indications of performance: 'all', 'next', and 'prev'. The 'all' AUC ROC measure represents the ability

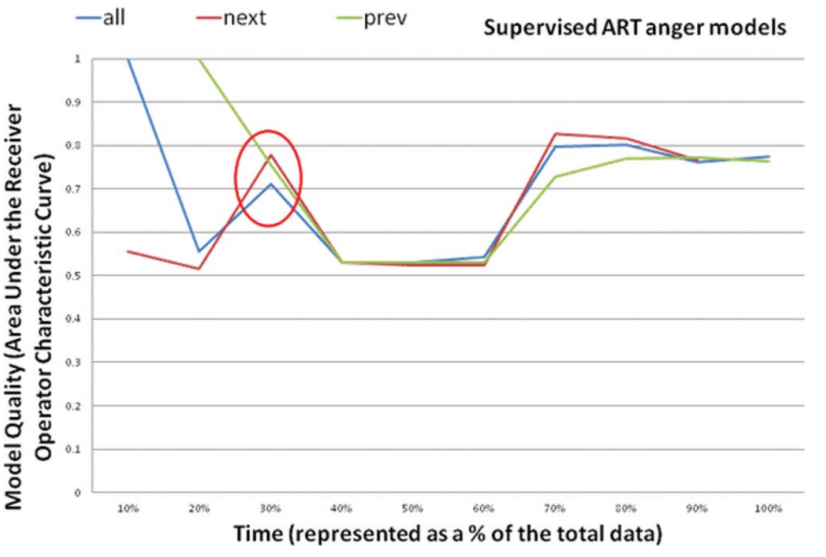

Figure 2. Measures of model quality (all/next/prev) for User 4137, shown to adjust in concert. Results typical across participants.

of the model to correctly classify all of the data that has so far been presented. As an example of the 'all' measure, the affective model of 50% of the data is compared to the true class labels of the 50% of the data that was presented. The 'prev' measure represents the ability of the current model to accurately classify the most recently observed data. 'Recently observed', in this instance, refers to the previous 10% of data. The 'next' measure represents the ability of the current model to accurately predict the upcoming class labels. 'Upcoming data', in this instance, refers to the next 10%. The measurements of these three items indicate whether a method is able to correctly model the data presented recently, in total, and/or in the near future. These three measures were shown to adjust in concert after a short time (30% of total time) as shown in Figure 2. This was found to be typical for all participants. Given that we are most interested in knowing the most recent affective state experienced by the learner, the most relevant feature is then 'previous' and this is what we selected for making our comparisons as well as for graphical and numerical representations in this paper.

The total data stream, of course, must be treated as though it has infinite length, as it would in a production environment. This type of time analysis was done to evaluate our research questions, as stated earlier, relating to (1) whether individualised affective models, created and utilised concurrently in real time during a tutoring session, perform better than their generalised and offline counterparts, and (2) determination of whether the model building process is positively supplemented by self-reported ground truth data in real time while the learner is being tutored.

4.1. Results: supervised real-time affective models

Models of Anger, Anxiety/Fear, and Boredom were created from the dataset and labels evaluated in Section 2, using only the supervised methods discussed in Section 3. Only

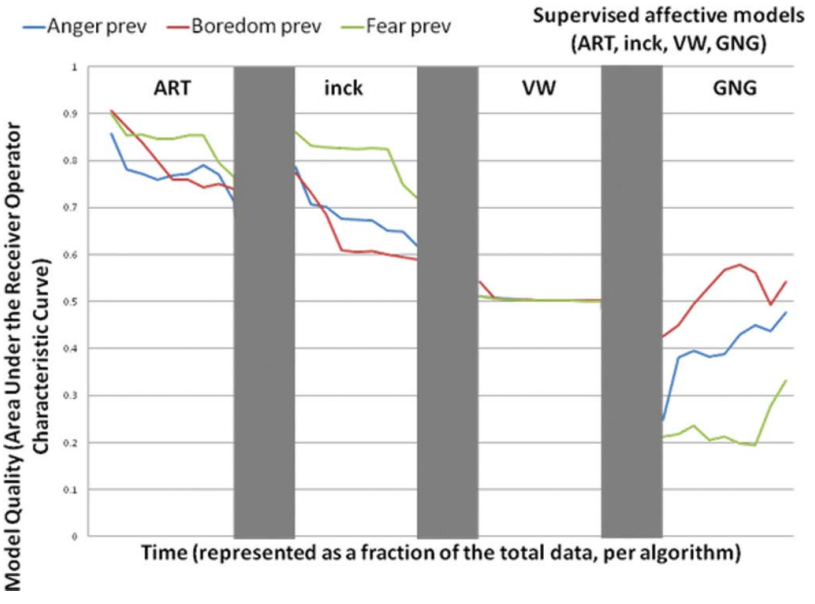

Figure 3. Affective modelling quality (*y*-axis, AUC ROC measure), as measured over time (*x*-axis, % of data measure) by AUC ROC on the most recent 10% of data, with all algorithms (four graphical segments) in supervised fashion.

supervised methods were used in order to enable an apples-to-apples comparison of real-time methods using labelled data to offline methods using labelled data. The results over time are shown in Figure 3.

Table 4 shows that acceptable *affective* models *are* able to be created in real time, as those models created with ART and with *k*-means remain above 0.6 AUC values for the majority of the time across all created models. All three affective models built with ART result in final model quality higher than 0.7. Two of the three clustering models (with the exception of Boredom and then only at the very end of the cycle) also result in comparable quality. However, from visual inspection of these figures, it is clearly evident that VW and GNG were at no point able to exceed the 0.6 AUC threshold of acceptability for any of the models. The complicated and dynamic nature of the provided graphs call for a more in-depth discussion of the two best-performing methods: ART and incremental *k*-means (abbreviated 'inck' in figures). A sample of the full results for this experiment is shown in Table 5. These results are displayed in summary because of page length limitations and the space-consuming nature of these tables. More complete numerical results are shown in (Brawner 2013). These results are shown in summary in Table 4, and complete Table 5.

Table 4. Summary of time-averaged supervised ART and clustering AUC when compared with offline equivalents.

Model	Anger	Anxiety/fear	Boredom
Offline	<0.6	0.83	0.79
ART	0.776	0.841	0.796
k-means	0.681	0.810	0.644

Table 5. Example Anger model qualities, supervised ART.

User	20%	30%	40%	50%	60%	70%	80%	90%	100%	Avg
4134	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.54	0.947
4133	0.58	0.58	0.58	0.58	0.54	0.51	0.68	0.69	0.50	0.584
4131	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.000
4127	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.000
4121	1.00	0.67	0.69	0.71	0.70	0.71	0.77	0.82	0.70	0.753
4111	0.63	0.81	0.79	0.78	0.79	0.80	0.79	0.66	0.74	0.756
4115	0.99	0.87	0.95	0.97	0.97	0.90	0.75	0.74	0.75	0.878
4135	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.000
4136	0.78	0.64	0.65	0.62	0.70	0.78	0.74	0.77	0.79	0.719
4137	1.00	0.76	0.53	0.53	0.53	0.73	0.77	0.77	0.76	0.709
4101	1.00	1.00	1.00	1.00	1.00	1.00	0.81	0.50	0.50	0.868
4117	0.56	0.52	0.50	0.50	0.57	0.53	0.58	0.56	0.56	0.545
4102	0.56	0.56	0.56	0.56	0.66	0.50	0.73	0.78	0.50	0.602
4105	0.76	0.70	0.76	0.66	0.65	0.64	0.70	0.70	0.58	0.682
4104	1.00	0.68	0.85	0.86	0.86	0.87	0.87	0.87	0.87	0.859
4107	1.00	1.00	0.99	0.63	0.63	0.63	0.63	0.63	0.63	0.749
4106	0.63	0.63	0.50	0.66	0.67	0.64	0.68	0.69	0.70	0.645
4112	0.91	0.64	0.64	0.67	0.58	0.69	0.76	0.77	0.84	0.723
4132	0.87	0.75	0.67	0.70	0.75	0.74	0.74	0.72	0.56	0.724
Average	0.857	0.780	0.772	0.760	0.768	0.772	0.790	0.771	0.712	0.776
Total Usable (avg ROC >0.6):				17		Percent usable:			89%	

Among the trends seen in the graphical data and tabular data is an overall *decrease* in AUC value as more data is obtained. This trend is generally observed in all sections, but is not unexpected. There are three main reasons for this downward trend.

First, as a byproduct of the AUC metric, the majority of the classifiers start with unreasonably high accuracy. As an example, all 10% of the initial data for User 4104 is of a single class, which results in a 100% accurate classifier. A general decrease in accuracy would be expected after having multiple classes to detect, for all algorithms.

Second, classification accuracy generally increases over time per user, but decreases slightly over time per group because of the larger magnitude of decrease when compared to increase. This is why average values per user are used to approximate overall model fit. User 4136 in Table 5 is an example of this expected tendency, as the reader can see a large decrement in classification accuracy brought on by initial class change, with gradual improvement further on. While other individual models have good performance throughout, the sharp initial drop of other models overwhelms the overall trend of the group average model performance over time. Lastly, the models do not have good fit overall for all users. An example of poor model fit can be seen in User 4117. This leads to the conclusion that such methods may not be appropriate for all users at all times. This phenomenon is discussed further in our conclusion of this paper, when considering how many of the models are usable (in addition to their overall quality). The group averaged, time averaged, values presented in Table 4 represent answers to the 'how useful, on average, would this approach be?'

4.2. Results: unsupervised real-time affective

If models of reasonable quality can be created *without* the use of labelled information, this would mark a significant improvement on the original offline models, as models of

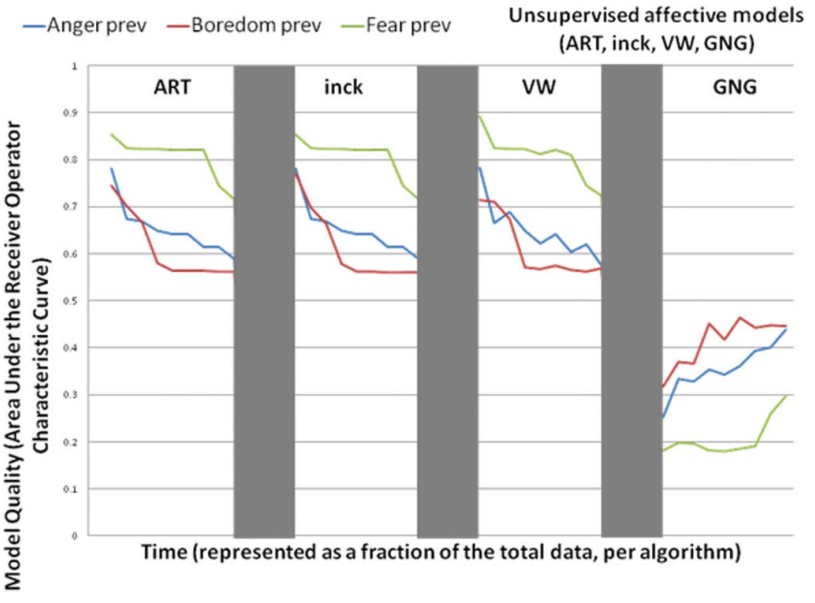

Figure 4. Affective modelling quality, as measured over time by AUC ROC on the most recent 10% of data; all algorithms unsupervised.

users could be created without their direct knowledge or interaction, aside from sensor measurement. Only unsupervised versions (without the benefit of labels) of the methods discussed in Section 3 were used in this section. The results over time are shown in Figure 4.

It is no surprise that, as seen in Figure 4, the unsupervised models perform more poorly in overall quality. The use of labelling information allows models to be of higher quality overall. These algorithms, however, were created for their use in real-world settings, where labelling information is not always available with fine resolution. There is no comparison against offline models for unsupervised models, as it is not appropriate to compare a model that includes labelled information with one that does not. The testing of unsupervised parameters allowed us to estimate how well constructed model quality can be when transferred to the field of use.

Nevertheless, the Anxiety/Fear models built with ART, *k*-means and VW performed well in spite of having no labels. The Anger models for these three algorithms, although registering inferior performance than Anxiety/Fear, also performed reasonably well throughout most of the cycle, slipping beneath the 0.6 AUC threshold only at the very end of the cycle. The Boredom models for these three algorithms started out well, but slipped below the 0.6 AUC minimal acceptance line (Section 2.3) line halfway through the cycle. All models built with GNG proved to behave unacceptably poorly.

Unsupervised models were built to represent the *worst possible performance,* as represented by creating models without labelled information. This sets the lower bound for comparison of the semi-supervised methods, which more closely approximate the real-world problem. This lower bound can be compared against the two established fully supervised bounds presented by offline and online approaches. A discussion of these results is found in the concluding section.

Table 6. Numerical summary of modelling results for ART and clustering.

	Total[1] or average[2] model quality				Individually usable datasets (of 19)			
Model	Anger	Boredom	Boredom (reduced)	Fear	Anger	Boredom	Boredom (reduced)	Fear
Offline Linear Regression[1]	< 0.6 (failure)	0.79*	0.79	0.83	*Cross-validated, but not expected to transition for reasons described in early sections*			
Supervised ART[2]	0.776	0.796	0.796	0.841	17	18	18	15
Unsupervised ART[2]	0.652	0.612	0.612	0.805	6	7	7	12
Semi-Supervised ART[2]	0.652	0.612	0.612	0.805	6	7	7	12
Supervised Clustering[2]	0.681	0.644	0.644	0.810	9	9	9	12
Unsupervised Clustering[2]	0.652	0.612	0.612	0.805	6	7	7	12
Semi-Supervised Clustering[2]	0.677	0.626	0.627	0.810	11	9	9	16

4.3. Results: semi-supervised real-time affective

As discussed in Section 1, it may be feasible on occasion to ask the user directly for a point of labelled data. Five labelled points (selected by the algorithms themselves) are used for each learner, which simulates directly asking the user about his/her affective state once every five minutes. For the models with barely acceptable average quality, does the injection of the occasional labels help?

We undertook a technique of *simulated* student querying to simulate the task of asking the user for their affective state. Roughly, 30% of the total labels are used in this process, and on-the-fly active learning is used algorithmically with the approach outlined in Section 3.5. Table 6 shows the effect that this has on overall model quality.

In short, the labelling information for a few points did not provide much additional value (if any) in affective state classification. The small number of label requests yields relatively little improvement when compared to zero requests for labels (unsupervised). This can be seen through the comparison of Figures 5 and 4. This small performance increase, however, for certain applications, may be the difference between acceptable and unacceptable model quality, as shown in detail as part of Table 4.

4.4. Results: discussion of VW and GNG

It became clear early on that these two algorithms were not suitable for this application. In this section we discuss why this may have been the case.

4.1.1. Growing neural gas

When data are closely aligned in the sampling space, segmentation of the data becomes difficult, and the GNG algorithm becomes more challenged. Each of the features in the dataset is raw, contains little to no preprocessing, and is not clearly segmentable over time. Additionally, the features have a tendency to move through the sampling space fluidly, leading to difficulty in the establishment of classification boundaries. These two features of the data determine the approach of the GNG algorithm on the problem, leading to a general trend that the GNG approach establishes *one large classification cluster* of the entire sampling space. This large cluster grows until it has encompassed all of the data

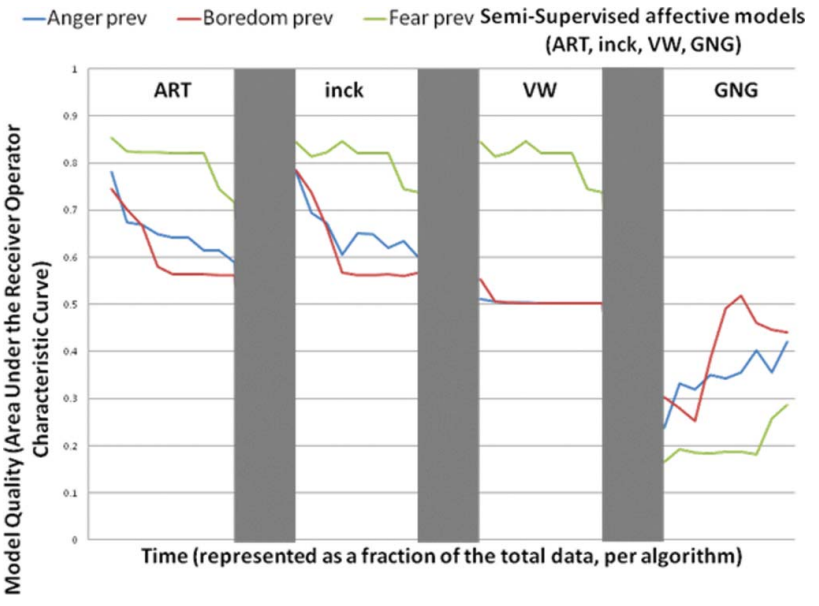

Figure 5. Affective modelling quality, as measured over time by AUC ROC on the most recent 10% of data, with all algorithms in semi-supervised fashion.

available, with few exceptions. The ROC measure for such a cluster is 0.5. While the GNG algorithm appears to 'improve' in quality over time and eventually reaches 0.5 AUC, it is a model of the baseline majority-class classifier, and does not produce usable models for any of the research questions.

This phenomenon is surprising and suggests a significant research gap. The OSSGNG model implemented by Beyer and Cimiano [51] is the only approach in this work that met all of the real-time algorithm checklist real-time-capable features discussed in Section 3. The observation that GNG does not produce usable models in any condition renders the safe removal of this approach from the discussion.

4.4.2. Vowpal Wabbit

Each algorithm models a different approach. While GNG represents a topographical overlay of the data, VW represents an incremental approach to linear regression modelling. VW adjusts weight vectors towards classes of labelled data, which increases reliance on labels. VW performs much better than the other algorithms when there are few states and feature sets to model.

New concept detection, however, has disastrous results in its overall performance. VW degrades to minimum performance quickly, and does not display any aptitude towards individual model recovery. The brittleness of the VW models is displayed through the above sections, with baseline-approaching performance. Although VW had an initially higher performance, when compared to the rest of the algorithms implemented in this work, it also had baseline performance for the longest period of time. Additionally, the VW models behave in more brittle fashion and benefitted the least from performance boost from labelling information.

Therefore, further discussion of GNG and VW has been omitted here in favour of discussion about ART and *k*-means clustering, as the overall performance of the last two has proven superior.

4.5. Results: conclusions for real-time models

As discussed earlier, the offline linear regression models created initially did not make use of all features of the data, denoted in the summary table (*). For completeness, the reduced feature dataset is tested on the affective models, in order to answer the research question, 'When eliminating features determined to be of little use during offline analysis, is overall quality also improved for affective real-time models?' The reduced feature set (e.g. using only the NeuroSky and Zephyr Heart sensors) was created based on the offline analysis which revealed that the other sensors did not contribute to the total model of Boredom. The results of this test are added in Table 6.

In brief, and based on our results, we can confidently conclude that quality affective models *can* be constructed using supervised, unsupervised, and semi-supervised approaches, where very infrequent semi-supervision information, at least in certain representations, can increase the number of usable models beyond the other approaches. The results from the creation of the affective models are encouraging. The previously created affective models achieved quality of <0.6, 0.83, and 0.79 (see Table 3), while supervised ART is able to *outperform*, on *all benchmarks*, the offline approach *using a small fraction of the total data*. This succinctly confirms that online models can be created, and indicates what future research in this area would improve these models further.

Our research described here has not lost track of the goal of creation and use of student models for use in an ITS setting. With this goal in mind, a more valuable metric of success is how well the algorithms for creating models perform when given little labelling information, as is the case within an ITS during a training session. When evaluating our research results by this metric, the unsupervised ART and unsupervised clustering models are equivalent, while the offline models are expected to have poor quality for the reasons discussed in Section 1. Our research here indicates that our approach to building individualised online models *would be expected to transfer* to use. To the knowledge of the authors, no other authors in the literature have made the claim that their affective models would expect successful transfer.

5. Summary

Each chosen method represents a different approach to establishing models from data in real time. Online clustering represents the method of dealing with online data of unknown classification through establishing and adjusting areas of the sampling space. Vowpal Wabbit represents the online approach to linear regression modelling, corresponding to the initial offline modelling approach chosen by the original experimenters. Adaptive Resonance Theory represents a neural network approach to online modelling, previously shown to have good one-pass learning results. GNGs represent the Self Organizing Map approach to establishing structure among data. The observation of performance of these real-time methods with physiological data shows that the approach is valid, and serves to recommend future solutions in the clustering and neural approaches.

This validity is shown in several ways. Firstly, direct comparison of fully supervised results indicates that real-time methods outperform their offline equivalents. Secondly, AUC values shown from the created models indicate that they are of acceptable overall quality. Such models are able to be created with less overall labelling information while exploiting user presence. Lastly, since such models are made from scratch, there is no reason to suspect that they suffer from cultural or population biases.

6. Future work

In the first sections of this work, we contend that learner models of affect and cognition can aid in the selection of a learning strategy, and that a learner model should be created using an individualised and real-time approach, rather than a general model created offline from data collected long time before model creation. We proceed to show that it is possible to build models for classification of affect in real time as they are needed that are individualised – that is, for a particular individual during the time of his/her tutoring session. The clearest avenue for future work is the integration of this work into an ITS.

This work is not without flaw, and requires further validation. The dataset used in this research met the majority of the specifications for an ideal dataset, but not all. A follow-on data collection and offline analysis effort is currently in progress with over 100 participants. This study will provide additional validation of methods in addition to expanding into interaction behaviours.

The methods presented here for real-time modelling were not created for the purpose of research and discovery; their potential use drives their development. The logical next step is to merge the work presented here into an intelligent tutoring system – for testing, validation, or for use. At the time of this writing, the Generalized Intelligent Framework for Tutoring (GIFT) project by Army Research Laboratory has over 600 users, several running experiments, and a recent workshop at the ITS conference. It is anticipated that the next version will incorporate the improvements reported here in individualised student modelling. The conclusions of this work will be presented to the community through integration into this community-driven research platform.

GIFT has been designed based on the idea of a learning effect chain (Sottilare et al. 2012). This has the derived requirement for separable software modules, which have defined interfaces. The defined process of the learner module is to take sensor and performance data and form it into a 'picture of the learner' from which to make pedagogical decisions. The current work is targeted to make these decisions.

Of course, knowledge of student state is not enough information, by itself, to inform how instruction should be adapted. For example, a learner which is anxious during test-taking may require no instructional intervention, while a leaner anxious during initial training exposure may need the pace of material presentation slowed. GIFT 3.0 presents a framework for pedagogy, as informed by state classification machines that adjusts content. Other work has been done to create domain-independent pedagogy (Goldberg et al. 2012), through an Engine for Macro-Adaptive Pedagogy. Further developments are currently in progress for a strategy recommendation engine for micro-adaption, which will likely be more state-dependent than its macro-adaptive counterpart.

This research to classify affective and cognitive states is intended to function as a part of architecture to support intelligent tutoring. The GIFT architecture is the intended

architecture for the transition of this technology. It collects various sensor characteristics such as electro-dermal response, and posture data from the Microsoft Kinect. It makes instructional strategy recommendations based on a decision tree of traits, states, and performance. It does not, however, contain a module for merging performance and sensor data into states into decisions. The work presented is the first of its kind to do so in a manner which can withstand validation, and presents a path for use.

Acknowledgements

The authors would like to thank Dr Bob Sottilare and the Army Research Laboratory for providing overwhelming support for this research. We also would like to thank Dr Benjamin Goldberg and Design Interactive for data collection.

Disclosure statement

No potential conflict of interest was reported by the authors.

About the authors

Keith W Brawner received his BS, MS, and PhD degrees from the University of Central Florida with a focus in Intelligent Systems and Machine Learning while working for the Department of Defense. He currently works for the Army Research Laboratory's Simulation and Training Technology Center in Orlando, Florida. He conducts research and manages research projects on the topic of Intelligent Tutoring Systems.

Avelino J Gonzalez is a professor in the Department of Computer Science at the University of Central Florida. His research interest is in areas of artificial intelligence, context-based behaviour and representation, machine learning from observation, and interactive virtual human technology.

References

AlZoubi, O., R. Calvo, and R. Stevens. 2009. "Classification of EEG for Affect Recognition: An Adaptive Approach." In *AI 2009: Advances in Artificial Intelligence*, 52–61. Heidelberg: Springer Berlin.

Alzoubi, O., S. Hussain, S. D'Mello, and R.A. Calvo. 2011. "Affective Modeling from Multichannel Physiology: Analysis of Day Differences." In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction (ACII 2011), LNCS*, edited by S. D'Mello, A. Graesser, B. Schuller and J-C. Martin, 4–13. Berlin: Springer-Verlag.

AlZoubi, O., I. Koprinska, and R.A. Calvo. 2008. "Classification of Brain-Computer Interface Data." In *Proceedings of the 7th Australasian Data Mining Conference*, 87: 123–131.

Dempsey, M.E. 2011. *The US Army Learning Concept for 2015* (TRADOC Pam 525-8-2). Washington, DC: Department of the Army HQ, US Training and Doctrine Command.

Baker, R.S., A.T. Corbett, K.R. Koedinger, and A.Z. Wagner. 2004. "Off-Task Behavior in the Cognitive Tutor Classroom: When Students Game the System." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 383–390. New York, NY: ACM.

Baker, R.S.J., J. Kalka, V. Aleven, L. Rossi, S.M. Gowda, A.Z. Wagner, G.W. Kusbit, M. Wixon, A. Salvi, and J. Ocumpaugh. 2012. *Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra*. Washington, DC: International Educational Data Mining Society.

Beringer, J., and E. Hüllermeier. 2006. "Online Clustering of Parallel Data Streams." *Data & Knowledge Engineering* 58 (2): 180–204.

Berka, C., D.J. Levendowski, M.N. Lumicao, A. Yau, G. Davis, V.T. Zivkovic, R.E. Olmstead, P.D. Tremoulet, and P.L. Craven. 2007. "EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks." *Aviation Space and Environmental Medicine* 78 (5): B231–B244.

Beyer, Oliver, and Phillip Cimiano. 2011. "Online Semi-Supervised Growing Neural Gas." Workshop New Challenges in Neural Computation 2011.

Beygelzimer, Alina, Daniel Hsu, John Langford, and Tong Zhang. 2010. "Agnostic Active Learning Without Constraints." In *Advances in Neural Information Processing Systems*, 199–207. New York, NY: ACM.

Blanchard, E., P. Chalfoun, and C. Frasson. 2007. "Towards Advanced Learner Modeling: Discussion on Quasi Real-Time Adaptation with Physiological Data." In *Proceedings of the 7th IEEE International Conference on Advanced Learning Technologies*, 809–813. Montreal, Quebec.

Bloom, B. S. 1984. "The 2-Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring." *Educational Researcher* 13 (6): 4–16.

Brawner, Keith W. 2013. "Modeling Learner Mood In Realtime Through Biosensors For Intelligent Tutoring Improvements." Doctor of Philosophy in EECS, Department of Electrical Engineering and Computer Science, University of Central Florida.

Brawner, Keith. 2014. *Data Sharing: Low-Cost Sensors for Affect and Cognition*. London: Educational Data Mining.

Brawner, Keith W., and A.J. Gonzalez. 2011. "Realtime Clustering of Unlabeled Sensory Data for User State Assessment." In *Proceedings of International Defense & Homeland Security Simulation Workshop of the I3M Conference*, Rome, Italy, September.

Calvo, Rafael A., and Sidney D'Mello. 2010. "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications." *IEEE Transactions on Affective Computing* 1 (1): 18–37.

Carpenter, G.A., and S. Grossberg. 1987. "A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine." *Computer Vision, Graphics, and Image Processing* 37 (1): 54–115.

Carpenter, G.A., and S. Grossberg. 1995. "Adaptive Resonance Theory (ART)." In *The Handbook of Brain Theory and Neural Networks*, edited by M. Arbib, 79–82. Cambridge, MA: MIT press.

Carroll, M., C. Kokini, R. Champney, R. Sottilare, and B. Goldberg. 2011. "Modeling Trainee Affective and Cognitive State Using Low Cost Sensors." In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*. Orlando, FL.

Champney, R.K., and K.M. Stanney. 2007. "Using Emotions in Usability." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 51 (17): 1044–1049.

Chaouachi, M., and C. Frasson. 2010. "Exploring the Relationship Between Learner EEG Mental Engagement and Affect." In *10th International Conference on Intelligent Tutoring Systems*, Pittsburgh, PA.

Conati, C. 2002. "Probabilistic Assessment of User's Emotions in Educational Games." *Journal of Applied Artificial Intelligence* 16: 555–575.

Conati, C., and H. Maclaren. 2004. "Evaluating a Probabilistic Model of Student Affect." In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems, Lecture Notes in Computer Science* edited by J.C. Lester, R.M. Vicari, and U. F. Paraguac, 55–66. Berlin: Springer-Verlag.

Cooper, D., K. Muldner, I. Arroyo, B. Woolf, and W. Burleson. 2010. "Ranking Feature Sets for Emotion Models Used in Classroom Based Intelligent Tutoring Systems." *User Modeling, Adaptation, and Personalization*, 135–146. Berlin Heidelberg: Springer.

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

D'Mello, S. K., R. Taylor, and A. C. Graesser. 2007. "Monitoring Affective Trajectories during Complex Learning." In *Proceedings of the 29th Annual Cognitive Science Society*, edited by D.S. McNamara and J.G. Trafton, 203–208. Austin, TX: Cognitive Science Society.

Fritzke, B. 1995. "A growing neural gas network learns topologies." *Advances in Neural Information Processing Systems* 7: 625–632.

García-Rodríguez, J, F Flórez-Revuelta, and JM García-Chamizo. 2007. "Image Compression Using Growing Neural Gas." *Neural Networks*, 2007. IJCNN 2007. International Joint Conference on.

Goldberg, Benjamin, Keith Brawner, Robert Sottilare, Ron Tarr, Deborah R. Billings, and Naomi Malone. 2012. "Use of Evidence-based Strategies to Enhance the Extensibility of Adaptive Tutoring Technologies." In *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*. Arlington, VA: National Training and Simulation Association.

Graesser, A., P. Chipman, B. King, B. McDaniel, and S. D'Mello. 2007. "Emotions and Learning with AutoTutor." In *Proceedings of the 13th International Conference on Artificial Intelligence in Education (AIED 2007)*, edited by R. Luckin, K. Koedinger and J. Greer, 569–571. Amsterdam: IOS Press.

Gross, R., I. Matthews, J. Cohn, T. Kanade, and S. Baker. 2010. "Multi-pie." *Image and Vision Computing* 28 (5): 807–813.

Hanley, J.A. 1989. "Receiver Operating Characteristic (ROC) Methodology: The State of the Art." *Critical Reviews in Diagnostic Imaging* 29 (3): 307.

Hanley, J.A., and B.J. McNeil. 1983. "A Method of Comparing the Areas Under Receiver Operating Characteristic Curves Derived from the Same Cases." *Radiology* 148 (3): 839–843.

Hewig, J., D. Hagemann, J. Seifert, M. Gollwitzer, E. Naumann, and D. Bartussek. 2005. "A Revised Film Set for the Induction of Basic Emotions." *Cognition and Emotion* 19 (7): 1095.

Holmstrom, J. 2002. "Growing Neural Gas." Master's thesis, Uppsala University.

Hulten, G., L. Spencer, and P. Domingos. 2001. "Mining Time-Changing Data Streams." In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 97–106. New York, NY: ACM.

Jain, A.K. 2008. "Data clustering: 50 years beyond k-means." In *ECML PKDD '08: Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I*, 3–4. Berlin: Springer-Verlag.

Jones, David, Kelly Hale, Sara Dechmerowski, and Hesham Fouad. 2012. "Creating Adaptive Emotional Experience During VE Training." In *Proceedings of The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*, 1–10. Arlington, VA: National Training and Simulation Association.

Kim, Y., and A. Baylor. 2006. "A Social-Cognitive Framework for Pedagogical Agents as Learning Companions." *Educational Technology Research and Development* 54 (6): 569–596.

Koedinger, Kenneth R., John R. Anderson, William H. Hadley, and Mary A. Mark. 1997. "Intelligent Tutoring Goes To School in the Big City." *International Journal of Artificial Intelligence in Education* 8: 30–43.

Kokini, C., M. Carroll, R. Ramirez-Padron, K. Hale, R. Sottilare, and B. Goldberg. 2012. "Quantification of Trainee Affective and Cognitive State in Real-Time." In *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*. Arlington, VA: National Training and Simulation Association.

Langford, J, L Li, and A Strehl. 2007. *Vowpal Wabbit Online Learning. Technical Report*.

Lepper, M., and M. Hodell. 1989. "Intrinsic Motivation in the Classroom." In *Research on Motivation in Education Vol. 3*, edited by C. Ames and R.E. Ames, 73–105. New York, NY: Academic Press.

McQuiggan, S., S. Lee, and J. Lester. 2007. "Early Prediction of Student Frustration." *Affective Computing and Intelligent Interaction* 2007: 698–709.

Mott, Bradford W., and James C. Lester. 2006. "Narrative-Centered Tutorial Planning for Inquiry-Based Learning Environments." *Intelligent Tutoring Systems*, 675–684. Berlin Heidelberg: Springer.

Prudent, Yann, and Abdellatif Ennaji. 2005. "An Incremental Growing Neural Gas Learns Topologies." *Neural Networks*, 2005. IJCNN'05. *Proceedings. 2005 IEEE International Joint Conference on*.

Rosasco, Lorenzo, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. 2004. "Are Loss Functions All the Same?" *Neural Computation* 16 (5): 1063–1076.

Sabourin, Jennifer, Bradford Mott, and James C. Lester. 2011. "Generalizing Models of Student Affect in Game-Based Learning Environments." In *Proceedings of the 4th International*

Conference on Affective Computing and Intelligent Interaction (ACII 2011), LNCS, edited by S. D' Mello, A. Graesser, B. Schuller and J-C. Martin, 588–597. Berlin: Springer-Verlag.

Sidney, K.D., S.D. Craig, B. Gholson, S. Franklin, R. Picard, and A.C. Graesser. 2005. "Integrating Affect Sensors in an Intelligent Tutoring System." In *Affective Interactions: The Computer in the Affective Loop Workshop at 2005 International Conference on Intelligent User Interfaces*, 7–13. Notre Dame, IN: University of Notre Dame.

Sottilare, Robert A. 2012. "Considerations in the Development of an Ontology for a Generalized Intelligent Framework for Tutoring." In *International Defense & Homeland Security Simulation Workshop Vienna*, Austria, September 2012.

Sottilare, Robert A., Keith W. Brawner, Benjamin S. Goldberg, and Heather A. Holden. 2012. "The Generalized Intelligent Framework for Tutoring (GIFT)." Retrived from gifttutoring.org.

VanLehn, Kurt. 2011. "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems." *Educational Psychologist* 46 (4): 197–221.

Verdú, E., L.M. Regueras, M. J. Verdú, J.P. De Castro, and M.A. Pérez. 2008. "Is Adaptive Learning Effective? A Review of the Research." In *Proceedings of the 7th WSEAS International Conference on Applied Computer & Applied Computational Science (ACACOS '08)*, edited by L. Qing, S. Y. Chen, A. Xu and M. Li, 710–715. Stevens Point, WI: WSEAS Press.

Woolf, B.P. 2010. *A Roadmap for Education Technology*. Retrieved from http://www.cra.org/ccc/docs/groe/GROE%20Roadmap%20for%20Education%20Technology%20Final%20Report.pdf.

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

Zaki, S.M., and H. Yin. 2008. "A Semi-Supervised Learning Algorithm for Growing Neural Gas in Face Recognition." *Journal of Mathematical Modelling and Algorithms* 7 (4): 425–435.

Zhu, X. 2005. "Semi-Supervised Learning Literature Survey." Technical Report 1530, Department of Computer Sciences, University of Wisconsin, Madison.