Towards An Affect-Sensitive AutoTutor

Sidney D'Mello[1], Rosalind Picard[2], and Arthur Graesser[1]

[1] The University of Memphis, {sdmello|a-graesser}@memphis.edu
[2] MIT Media Laboratory, picard@media.mit.edu

Abstract
This paper investigates the reliability of detecting a learner's affective states in an attempt to augment an Intelligent Tutoring System (AutoTutor) with the ability to incorporate such states into its pedagogical strategies to improve learning. We describe two studies that used observational and emote-aloud protocols in order to identify the affective states that learners experience while interacting with AutoTutor. In a third study, training and validation data were collected from three sensors in a learning session with AutoTutor, after which the affective states of the learner were identified by the learner, a peer, and two trained judges. The third study assessed the reliability of automatic detection of boredom, confusion, delight, flow, and frustration (versus the neutral baseline) from sensors that monitored the manner in which learners communicate affect through conversational cues, gross body language, and facial expressions. Although the primary focus of this article is on the classification of learner affect, we also explore how an affect-sensitive AutoTutor can adapt its instructional strategies to promote learning.

1. Introduction

Emotions (affective states) are inextricably bound to the learning process in addition to the well-documented impact of cognition, motivation, discourse, action, and the environment. Attempts to master difficult technical material, such as conceptual physics or mathematics, inevitably require learners to confront contradictions, anomalous events, obstacles to goals, salient contrasts, and other stimuli or experiences that fail to match expectations. In response to these discrepant events, the autonomic nervous system increases its arousal and the learner experiences emotions such as confusion, frustration, irritation, anger, rage, or even despair. Cognitive equilibrium is subsequently restored when discrepancies are resolved, misconceptions are discarded, and confusion is alleviated. At that point the learner resumes with hope, determination, renewed curiosity, and maybe even enthusiasm. Given this link between affect and cognition, an agile learning environment that is sensitive to a learner's affective states will presumably enrich learning, particularly when deep learning is accompanied by confusion, frustration, boredom, interest, excitement, and insight.

In this article we consider the possibility of endowing an existing Intelligent Tutoring System (ITS), AutoTutor, with the ability to process the learners' affective states in addition to their cognitive states. AutoTutor would ideally identify the learners' emotions and adjust its pedagogical strategies during the learning of complex material. AutoTutor is a fully automated computer tutor that simulates human tutors and holds conversations with students in natural language[1]. AutoTutor helps students learn Newtonian physics and computer literacy by presenting challenging problems (or questions) from a curriculum script and engaging in a mixed-initiative dialog while the learner and AutoTutor collaboratively construct an answer. AutoTutor provides feedback to the student on what the student types in (positive, neutral, negative feedback), pumps the student for more information ("What else?"), prompts the student to fill in missing words, gives hints, fills in missing information with assertions, identifies and corrects erroneous ideas, answers the student's questions, and summarizes topics. A full answer to a question is eventually constructed during this dialog, which normally takes between 30 and 100 student and tutor turns.

The underlying assumption behind the endeavor to develop the affect-sensitive AutoTutor is that affect is inextricably bound to learning. We are not alone in this view. Over the last few years there have been sustained efforts to incorporate assessments of the learner's affect into the pedagogical strategies of ITSs. Kort, Reilly, and Picard[2] proposed a comprehensive four-quadrant model that explicitly links learning and affective states. This model was used in the MIT group's work on their *affective learning companion*, a fully automated computer program that recognizes a learner's affect by monitoring facial features, posture patterns, and onscreen keyboard/mouse behaviors. Conati[3] has developed a probabilistic system that can reliably track multiple emotions of the learner during interactions with an educational game. Her system relies on dynamic decision networks to assess the affective states of joy, distress, admiration, and reproach. Litman and Silliman's work with their ITSPOKE[4] conceptual physics ITS has used a combination of discourse markers and acoustic-prosodic cues to detect and respond to a learner's affective states.

The achievement of an affect-sensitive tutorial interaction engages the tutor and learner in an *affective loop*. This loop includes the *identification* of the affective states relevant to learning, real-time *detection* of those states, the *selection* of appropriate tutor actions that maximize learning while influencing the learner's affect, and the *synthesis* of emotional expressions by the tutor as it attempts to engage the learner in a more human-like, naturalistic manner. The achievement of an affective loop in an integrated system can be viewed from the perspective of either the learner or the tutor. The learner-centric view consists of analyzing the prominent affective states in the learner, assessing their potential impact on learning, identifying how these states are manifested in the learner, and developing an automatic affect detection system. The tutor-centric view explores how good human tutors or theoretical ideal tutors adapt their instructional agenda to encompass the emotions of the learner. This expert knowledge is then transferred to computer tutors such as AutoTutor. Embodied conversational agents that simulate human tutors are programmed to synthesize affective elements through the generation of facial expressions, the inflection of speech, and the modulation of posture.

The development of the affect-sensitive AutoTutor has been guided by research in a number of fronts that span computer science, artificial intelligence, psychology, and the learning sciences. There are a number of differences between our approach and previous work. First, we consider a larger set of affective states (N = 7) than some of the earlier systems which have concentrated on intensity, valence, or small subsets of the affective states. This is important because a larger set of affective states is required to encompass the gamut of learning. A person's reaction to the presented material can change depending on their goals, preferences, expectations and knowledge state[3]. The second major difference involves our use of multiple human judges to measure affect. This is a significant step because previous empirical research has documented that humans are not particularly good at judging affect. This is potentially problematic when accurate models of ground truth are needed for supervised machine learning algorithms to classify affect. Our research involves the learner, a peer, and 2 trained judges in identifying the affective states of a learner in order to obtain a reasonable degree of convergent validity. The third major difference that distinguishes our approach from earlier research involves the use of relatively unique sensors for detecting affect. Over the last decade, researchers have achieved significant contributions towards the development of automated affect sensing systems (see Pantic, 2003[5]). These include systems that detect affect from bodily measures such as physiological signals (e.g., electromyography, heart rate monitors, skin conductance), facial features, and acoustic-prosodic features. We similarly track facial features but we also explore two relatively unexplored channels for affect detection. These include dialogue features extracted from the tutorial session and body posture.

The next section presents results from two studies that attempted to identify the affective states that learners typically experience while interacting with AutoTutor. We subsequently describe a study in which diagnostic data were collected with three sensors during a typical tutorial session. This study also investigated the reliability in which human judges recognized learners' affect states. Classification experiments were conducted that assess the reliability by which the affective states of a learner can be automatically recognized by a computer. We conclude by discussing the prospects of incorporating the learner's affect into AutoTutor's pedagogical strategies.

2. The Relationship between Affect and Learning

Learning inevitably involves failure and a host of associated affective responses. The vast majority of work in affective computing has focused on the 6 *basic* emotions that are ubiquitous in everyday experience. These include fear, anger, happiness, sadness, disgust, and surprise[6]. However, many have called into question the adequacy of basing an entire theory of emotions on these basic emotions[2]. Learners rarely experience sadness, fear, or disgust, for example. Therefore, the studies described below attempted to identify the affective states that learners typically experience while interacting with AutoTutor, with the expectation that these findings will generalize to other learning environments.

2.1. Observing Learner Emotions during Interactions with AutoTutor

Five trained judges observed different affect states (boredom, confusion, frustration, eureka, flow/engagement, versus neutral) that potentially occur during the process of learning introductory computer literacy with AutoTutor[7]. The participants were 34 college students. Trained judges recorded emotions that learners were apparently experiencing every 5 minutes during the 30-45 minute interaction with AutoTutor.

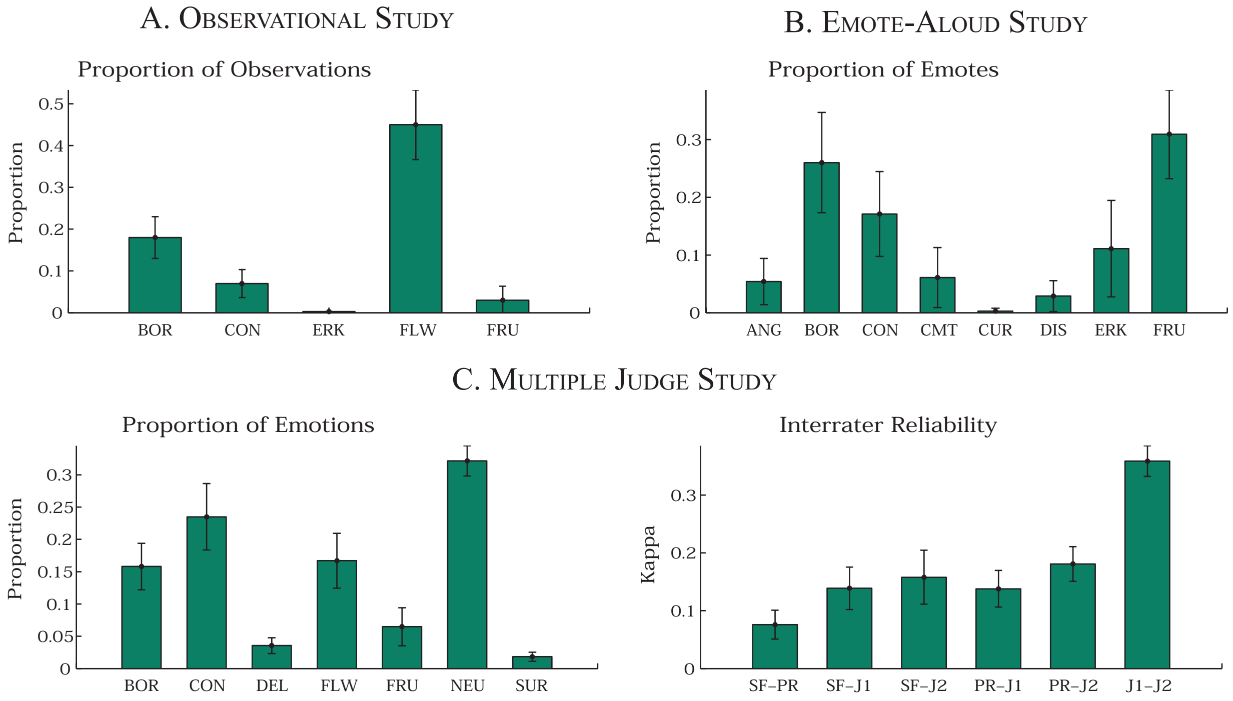
Figure 1a shows descriptive statistics on the proportions of the emotions observed in the tutoring sessions. The results revealed that experiences of eureka were much too rare in the experiment; there was only one recorded eureka experience in 17 total hours of tutoring among the 34 students. Frustration was also rarely experienced (only

3% of the recorded emotions). One explanation of the rare occurrence of frustration is that there were no practical consequences of poor performance in these experiments. We would expect frustration to be more prevalent when the learning experience is linked to student interest, to tasks the learner is vested in, or to high stakes testing. The percentage scores were higher for the affect states of confusion (7%), boredom (18%), and flow (45%).

2.2. Emote-Aloud while Students Interact with AutoTutor

In an emote-aloud procedure, college students (N = 7) verbalized their affective states while interacting with AutoTutor[8]. The affect states investigated were anger, boredom, confusion, contempt, curiosity, disgust, eureka, flow, and frustration. These affective states were defined for the learners before the experiment started.

We found that boredom, confusion, and frustration were reported at higher rates than were anger, contempt, and disgust, $F(7, 42) = 7.89$ ($p < .05$ in this all subsequent tests). This result confirms the hypothesis that Ekman's basic emotions play nonsignificant roles in learning, Although eureka was relatively well reported, we concluded that this response functionally signified delight from giving a correct answer rather than a deep eureka experience. It also appears that frustration was reported at a higher rate when participants verbalized their own emotions, as opposed to trained judges determining the emotions of the learner (Figure 1a). One explanation of this result may lie in the social display rules that people adhere to in expressive affect. Social pressures typically result in people disguising negative emotions, such as frustration, thus making it difficult for judges to detect this emotion. In contrast, when encouraged to freely reflect and report on their affect, as in the emote-aloud study, such barriers drop and frustration is freely expressed.



ANGer, **BOR**edom, **CON**fusion, Conte**M**p**T**, **CUR**ious, **DEL**ight,
DISgust,**EUR**eka, **FLO**w, **FRU**stration, **NEU**tral, **SUR**prise

SF: Self, **PR**: Peer,
J1: Trained Judge 1, **J2**: Trained Judge **2**

Figure 1: Proportions of affective states observed across 3 studies.

2.3. Measurement of Emotions by Multiple Judges

The training and testing of an emotion classifier needs a gold standard for comparison, i.e., some measure of ground truth of the affect of a learner. One approach to establishing a gold standard is to examine the reliability of humans in classifying emotions. We investigated three potential human-based measures of ground truth for emotion detection: the participants, novice judges, and trained judges[9].

We conducted a study in which college students (N = 28) interacted with AutoTutor for 32 minutes on topics in computer literacy (e.g., hardware, internet, or operating systems). Three streams of information were recorded during the participant's interaction with AutoTutor. A video of the participant's face was captured using the IBM®

blue-eyes camera. Posture patterns were captured by the Tekscan® Body Pressure Measurement System. A screen-capturing software program called Camtasia Studio was used to capture the audio and video of the participant's entire tutoring session. Figure 2 illustrates the various sensors utilized in the study.

The judging process was initiated by synchronizing the video streams from the screen and the face and displaying this to the judge. Judges were instructed to make judgments on what affective states were present at 20-second intervals; at each of these points, the video automatically paused (freeze-framed). Judges were also instructed to indicate any affective states that were present in between the 20-second stops.

Four sets of emotion judgments were made for the observed affective states of each participant's session with AutoTutor. For the self judgments, the participants watched their own session with AutoTutor immediately after completing the interaction. For the peer judgments, participants returned approximately a week later to watch and judge another participant's session on the same topic in computer literacy. Finally, two trained judges judged all of the sessions independently. These judges had been trained on AutoTutor's dialogue characteristics and also in the detection of facial actions according to Ekman's Facial Action Coding System (FACS[6]).

Figure 1c, (lower left) presents the proportion of judgments that were made for each of the affect categories, averaging over the 4 judges. The most common affective state was neutral (.32), followed by confusion (.24), flow (.17), and boredom (.16). The frequency of occurrence of the remaining states of delight, frustration and surprise were significantly lower, comprising .06, .04, and .02 of the observations respectively. This distribution of affective states implies that most of the time learners are either in a neutral state or in a subtle affective state (boredom or flow). There is also a reasonable amount of confusion. The incidence of confusion can be explained by the fact that the participants were typically low domain knowledge students, as indicated by their low pretest scores.

The design of this study allowed us to inspect and compare emotion judgments of the self, peer, and the 2 trained judges. Cohen's kappa scores, an index of interjudge reliability, were computed separately for each of the 28 learners; mean scores are presented in Figure 1c (lower right). Statistical analyses on the kappa scores revealed that there were significant differences among the six pairs, $F(5, 135) = 33.34$. The self-peer pair had the lowest inter-judge reliability scores when compared to the other five pairs. The two trained judges had significantly higher kappa scores than the other five pairs. These results support the conclusion that training on Ekman's Facial Action Coding System and tutorial dialogue can enhance the reliability and accuracy of judgments of affective states.

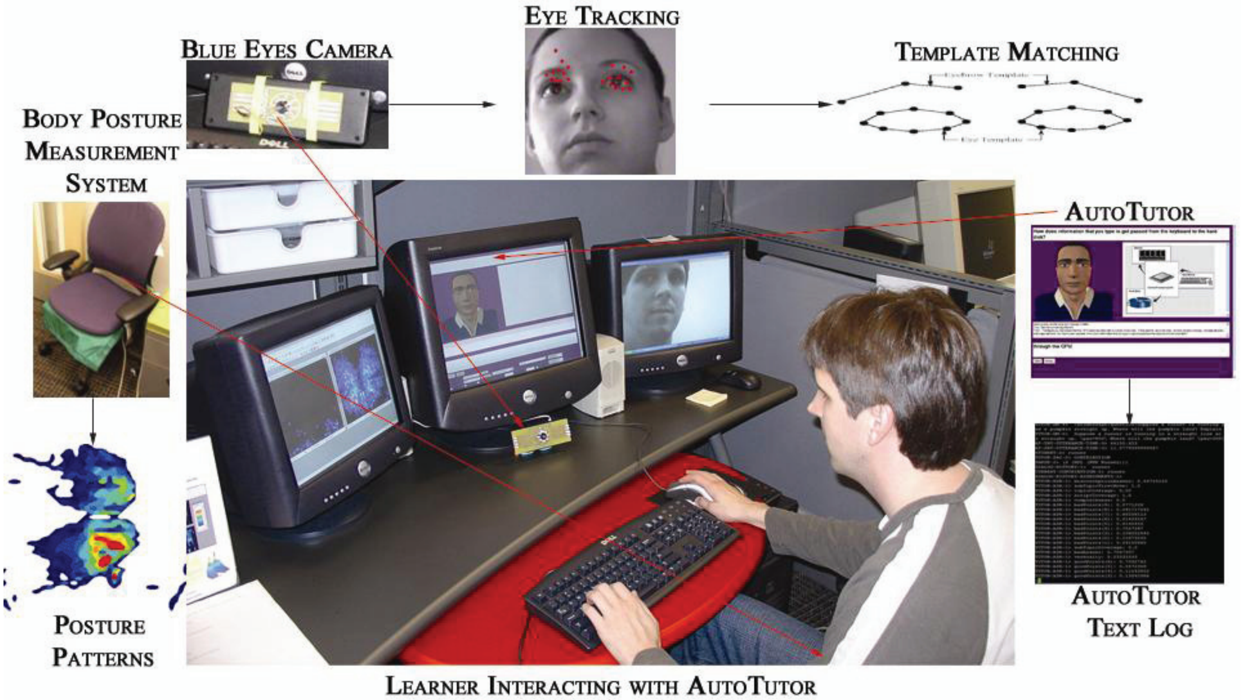


Figure 2: Sensors used in the multiple judge study.

3. Affect Detection from Conversational Cues

One of the advantages of affect detection in integrated learning environments is that a tutoring session supplies a rich trace of the interaction history. Several conversational features and discourse markers (collectively called dialogue features) can then be extracted and used to infer the learner's affect. The use of dialogue to detect affect in learning environments is a reasonable information source to explore because dialogue information is abundant in virtually all conversations and is inexpensive to collect.

3.1. AutoTutor Dialogue Features

A session with AutoTutor consists of a set of subtopics (main questions, e.g. How is the operating system loaded onto RAM?) that cover specific areas of the main topics (hardware, internet, and operating systems). Each subtopic has an associated set of expectations, potential dialogue moves to elicit expectations, corrections of misconceptions, and other slots in a *curriculum script* that need not be addressed here. The expectations are ideally covered over a series of turns in AutoTutor's conversation as the student attempts to construct an answer to the subtopic question. When an acceptable answer, with the appropriate details, is gleaned from the student's responses, AutoTutor moves on to the next subtopic. At the end of each student turn, AutoTutor maintains a log file that captures the student's response, assessments of the conceptual quality of the response, the feedback provided, and the tutor's next move.

We mined several features from AutoTutor's log files in order to explore the links between the dialogue features and the affective states of the learners. These features included temporal assessments for each student-tutor turn, such as the *subtopic number*, the *turn number* within a subtopic, and the student's *reaction time* (interval between presentation of the question and the submission of the student's answer). Assessments of response verbosity included the *number of characters* (letters, numbers) and *speech act* (that is, whether the student's speech act was a contribution towards an answer (coded as a 1) versus a frozen expression, e.g., "I don't know", "Uh huh" (coded as -1). The conceptual quality of the student's response was evaluated by Latent Semantic Analysis (LSA, http://lsa.colorado.edu/). LSA is a statistical technique that measures the conceptual similarity of two texts. LSA-based measures included a *local good score* (the conceptual similarity between the student's current response and the set of expectations being covered) and a *global good score* (the similarity of a set of student responses to a set of expectations in a good answer). Changes in these measures when compared to the previous turn were also included as the *delta local good score* and the *delta global good* score. AutoTutor's major dialogue moves were ordered onto a scale of conversational *directness,* ranging from -1 to 1, in terms of the amount of information the tutor explicitly provides the student: summary > assertion > prompt > hint > pump. AutoTutor's short *feedback* (negative, neutral negative, neutral, neutral positive, positive) is manifested in its verbal content, intonation, and a host of other non-verbal cues. The feedback was aligned on a scale ranging from -1 (negative feedback) to 1 (positive feedback).

Dialogue features were extracted for each turn and compared to the emotion judgments of the 4 judges. More specifically, the emotion judgment that immediately followed a dialogue move (within a 15 second interval) was bound to that dialogue move. This allowed us to obtain four sets of labeled dialogue data aggregated across the 28 participants. The sizes of these data sets were 1024, 1040, 1115, and 1119 for affect labels provided by the self, the peer, trained judge 1, and trained judge 2, respectively.

3.2. Relationship between Dialogue and Affect

We investigated the potential of dialogue features to discriminate particular emotions from the baseline of neural. Figure 3a shows descriptive statistics for dialogue features associated with an emotion versus neutral state. One-way ANOVAs were conducted on each of the dialogue features in order to explore significant relationships among the affect states. The results indicate that students experience more emotional episodes later in the session, as indicated by higher subtopic numbers, and also when they take longer to respond to questions posed by the tutor (higher response time). Emotional episodes occur with negatively coded speech acts, i.e. when students provide frozen expressions such as "I don't know" or "What?", as opposed to substantive answers to the tutor's queries. Our results also indicate that students experienced more emotions when they received negative feedback.

3.3. Automatic Detection of Affect from Dialogue

The Waikato Environment for Knowledge Analysis (WEKA) was used to evaluate the performance of various standard classification techniques in an attempt to detect affect from dialogue. The classifiers were selected to span a broad range of schemes such as Bayesian classification (Naïve Bayes classifier), neural networks (multilayer perceptron), functions (simple logistic regression), lazy classifiers (nearest neighbor), decision tree classifiers (C4.5 decision tress), and meta classification schemes (additive logistic regression). The classification algorithms were compared in their ability to detect boredom, confusion, delight, flow, and frustration from neutral. Surprise was excluded due to a very low number of observations. To establish a uniform baseline, we randomly sampled an equal number of observations from each affective state category. This process was repeated for 10 iterations and all reported reliability statistics were averaged across these 10 iterations. Classification reliability was evaluated on the 6 classification algorithms using k-fold cross-validation (k = 10).

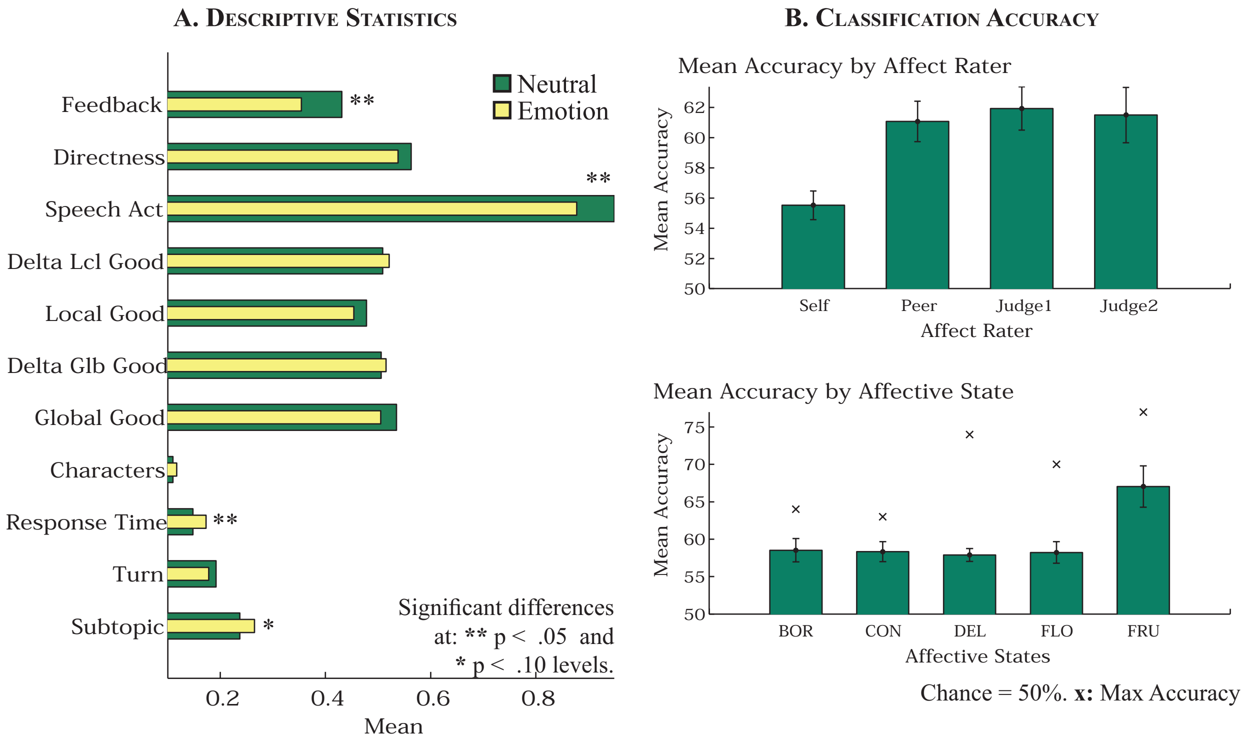


Figure 3: Descriptive statistics and reliability in affect detection from dialogue. For visualization purposes all features in 3a have been normalized to a 0-1 range.

Our results indicated that there were significant differences in classification accuracies among each of the 4 data sets, *F*(3,15) = 120.64. The classifiers trained and tested on datasets where affect judgements were provided by the peer and the 2 trained judges were significantly higher than those provided by the self (see top of Figure 3b). There were significant differences in classification accuracies among the various emotions, *F*(4,20) = 42.73. Frustration was detected with higher accuracy than boredom, confusion, delight, and flow (see bottom of Figure 3b).

The statistical results described above considered the reliabilities of all 6 classifiers in order to estimate the overall trend. However, from an engineering perspective, we are also concerned with the classifiers that yielded the best performance. Our results indicate that a simple logistic regression achieved the highest accuracies of 64%, 63%, 74%, and 70% in detecting boredom, confusion, delight, and flow from neutral. For frustration an optimal accuracy of 77% was obtained by C4.5 decision trees. These results support the hypothesis that dialogue features can be a reasonable source to measure the affective states that a learner is experiencing.

4. Affect Detection from Posture

The Body Posture Measurement System (BPMS), developed by Tekscan™, was used to monitor the gross body language of a student during a session with AutoTutor. The BPMS consists of a thin-film pressure pad (or mat) that can be mounted on a variety of surfaces. The pad is paper thin with a rectangular grid of sensing elements that provide a pressure reading in mmHg. The setup used in the multiple annotator study involved the use of one sensing pad placed on the seat of a Steelcase™ Leap Chair and another placed on the back of the chair. The output of the BPMS system consisted of two 38x41 matrices (for the back and seat) with each cell in the matrix corresponding to the amount of pressure exerted on the corresponding element in the sensor grid.

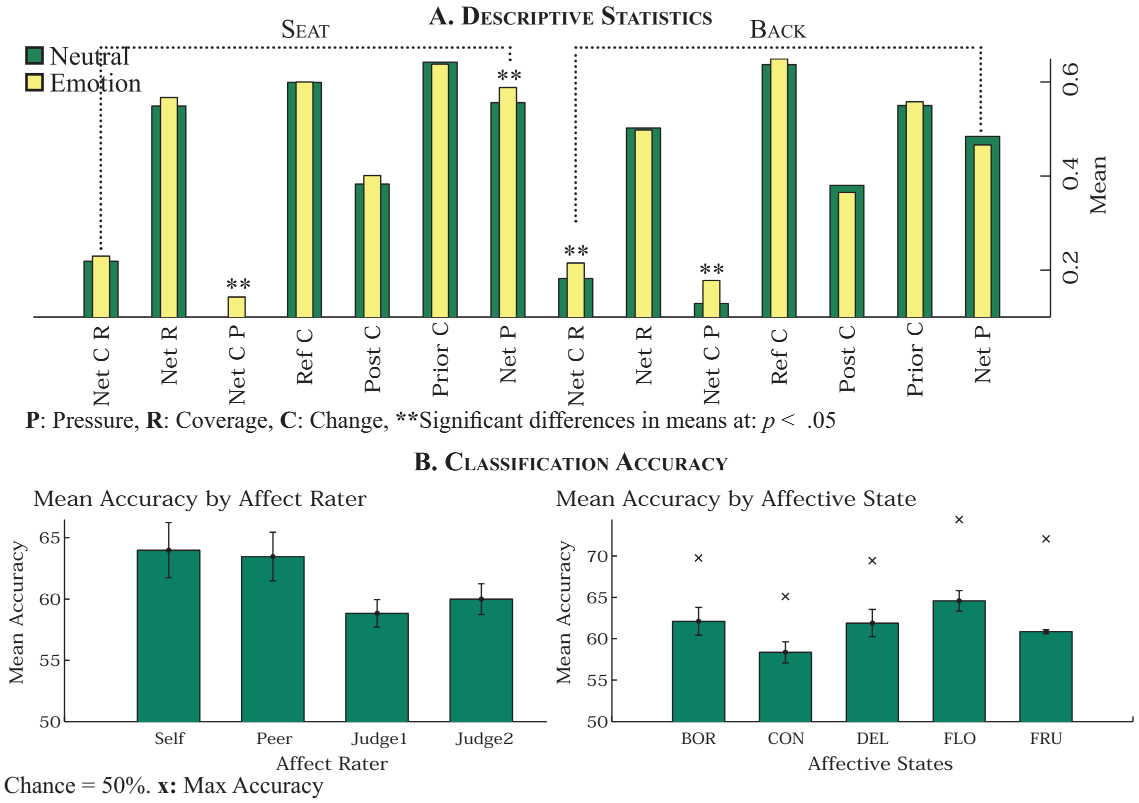
4.1. Posture Features

Several features were computed by analyzing the pressure maps of the 28 participants recorded in the study. We computed 5 pressure-related features and 2 features related to the pressure coverage for both the back and the seat, yielding 14 features in all. Each of the features was computed by examining the pressure map during an emotional episode (called the *current frame*). The pressure related features include the *net pressure,* which measures the average pressure exerted. The *prior change* and *post change* measure the difference between the net pressure in the current frame and the frame three seconds earlier and later respectively. The *reference change* measures the difference between the net pressure in the current frame and the frame for the last known affective rating. Finally, the *net pressure change* measures the mean change in the net pressure across a predefined window, typically 4 seconds, that spans two seconds before and two seconds after an emotion judgment. The two coverage features examined the proportion of non-negative sensing units (*net coverage*) on each pad along with the mean change of this feature across a 4-second window (*net coverage change*).

We created four data sets that temporally integrated the posture feature vectors with affect ratings provided by the four human judges. The sizes of these data sets were 2642, 2702, 3452, and 3377 for affect labels provided by the self, the peer, trained judge 1, and trained judge 2, respectively.

4.2. Relationship between Posture and Affect

Figure 4a shows descriptive statistics for each posture feature, segregated by emotion. One-way ANOVAs were conducted on each of the posture features in order to explore relationships with the emotions. Emotion episodes were distinguishable from neutral primarily with respect to a heightened increase in activity, manifested by large changes in pressure, on the back and the seat. These results confirm the hypothesis that affective states are typically accompanied by a degree of physiological arousal. The arousal is detected by the posture monitoring system. It should be noted that boredom is accompanied by the aroused state of fidgeting, not inactivity. These results also replicate the spirit of earlier findings by Mota and Picard at MIT, where they monitored activity related posture features and discovered that children fidget when they were bored while performing a learning task on a computer.


A. Descriptive Statistics

P: Pressure, **R**: Coverage, **C**: Change, ****Significant differences in means at: *p* < .05

B. Classification Accuracy

Chance = 50%. **x:** Max Accuracy

Figure 4: Descriptive statistics and reliability in affect detection from posture.

4.3. Automatic Detection of Affect from Posture

We performed affect classification analyses with posture, just as we previously reported for dialogue features. Standard classifiers were trained and tested on 4 data sets. each consisting of the 14 posture features and affect labels provided by the self, peer, and 2 trained judges. We found significant differences in classification accuracy across the 4 data sets, $F(3,15) = 25.30$. Classification accuracy for novice judges (self and peer) outperformed the trained judges (see left of Figure 4b). The classifiers were more successful in detecting flow (versus neutral) than they were for detecting boredom, confusion, and frustration, $F(4, 20) = 30.00$. Boredom and delight were more readily detected than confusion (see right of Figure 4b).

The best classifier for detecting each particular emotion versus neutral was the k-nearest neighbor classifier (k=1). This classifier achieved accuracies of 70%, 65%, 74%, and 72% in detecting boredom, confusion, flow, and frustration versus the neutral baseline. For delight, a logistic regression classifier had the best accuracy of 70%. These results confirm that posture can be a viable channel in inferring a learner's affect. Indeed, the overall classification accuracy of emotions based on posture was 70%, the same percentage achieved from dialogue.

5. Affect Detection from Facial Features

Ekman and Friesen[5] highlighted the expressive aspects of emotions with their Facial Action Coding System. This system specifies how "basic" emotions can be identified on the basis of facial behaviors and the muscles that produce them. Each movement in the face is called an *action unit* (or AU). There are approximately 58 action units altogether. These prototypical facial patterns have been used to identify the 6 basic emotions: happiness, sadness, surprise, disgust, anger, and fear. The development of a system that automatically detects the action units is quite a challenging task, however, because the coding system was originally created for static pictures rather than on changing expressions over time.

We are currently exploring some of the technical challenges associated with the automated detection of facial expressions. As an initial step, we have had two trained judges code a sample of the observations of emotions on the action units. The sample of coded affective states consisted of *voluntary* judgments in which both of the 2 trained judges agreed on the learner's emotion. Recall that judges were required to provide affect judgements every 20 seconds. Voluntary judgements include points in between those 20-second time spans where judges offered emotion judgements. The samples were selected to approximate an equal distribution of emotions from the 28 participants. The database consisted of 212 samples that included the affective states of boredom, confusion, delight, frustration, and neutral. Flow was excluded because it rarely occurred at the voluntary points. The two trained judges coded each of the samples separately and achieved a fair kappa score of .72 in detecting a subset of the action units.

A database of action units associated with emotions was created by considering the AU codings of both judges. The 6 standard classifiers were used to detect the affective states of boredom, confusion, delight, and frustration from neutral on the basis of the human coded AUs. The classification accuracy for delight was the highest (90%), boredom the lowest (60%), whereas confusion (76%) and frustration (74%) were in between. The classifiers were more successful in detecting emotions that are manifested with highly animated facial activity, such as delight, than emotions that are more subtly expressed (boredom). Additionally, the lower classification accuracies associated with frustration might be attributed to participants attempting to disguise negative emotions. Overall, the classification accuracy for facial expressions was a bit higher (75%) than those for body posture and dialogue (70% each).

While the facial features seemed to be good predictors of affect, the classification results should be interpreted with caution. This is because human judges annotated the facial action units of the learner, and we would expect some reduction in accuracy when the AUs are coded by a computer. However, recent results by el Kaliouby and Robinson[10] indicate that low reliabilities in the automated measurement of AUs may be compensated by more robust emotion classifiers such as Dynamic Bayesian Networks.

6. Discussion

This research was motivated by the belief that ITSs can be more than mere cognitive machines. We believe they can be endowed with the ability to recognize, assess, and react to a learner's affective state. We conducted two exploratory studies to explore the links between learning and emotions, links that have not yet been systematically tested in the ITS community. Our results supported the hypothesis that deep learning of conceptual material is dominated by boredom, confusion, delight, flow, and frustration rather than Ekman's basic emotions of anger, fear, happiness, sadness, disgust, and surprise.

The study with multiple judges allowed us to collect diagnostic data from the sensors as well as multiple perspectives on the learners' affective states. The design of our study provided an ecologically valid environment to monitor the natural emotions of a learner during a tutorial session with AutoTutor. This is a noticeable departure from several earlier studies on emotion that typically recruit actors and that intentionally induce emotions in a contrived context.

Although the problem of automating affect recognition is extremely challenging, on par with automating speech recognition, we achieved several milestones that suggest that significant information can be achieved in an automated way. In particular, we were able to achieve reasonable classification accuracies in detecting the various affective states from neutral. There were two relatively novel sensors (conversational dialogue and posture) and one traditional sensor (facial features). It is also important to note that the sensors used were non-invasive in the sense that they provided online measurement with minimal task interference; this is an important requirement for learning environments. Our classification results indicate that the posture sensor would be the sensor of choice for affective states that do not generate overly expressive facial expressions, such as boredom (70%) and flow (74%). On the other hand, the affective states of confusion (76%) and delight (90%), which are accompanied by significant arousal, are best detected by monitoring facial features. The negative affective state of frustration is typically disguised and therefore difficult to detect with the bodily measures of face and posture. Frustration is best detected by examining the dialogue features in the tutoring context (77%). Taken together, detection accuracies are over 77% when particular emotions are aligned with the optimal sensor channels.

Critics may object to our claim that the reported classification results have *reasonable* accuracy. However, we would argue that an upper bound on automated classification accuracy for affect has yet to be established and that there are no theoretical or empirical foundations for expecting it to be extremely high. Human classifications may be proposed as the ultimate upper bound on system performance. If so, available results suggest that human classification performance is hardly impressive[9] and not necessarily an improvement over machine classification. For example, el Kaliouby and Robinson[10] reported modest performance when a group of 18 people were asked to classify six affective states from a set of test videos. Humans had 54.5% accuracy scores, whereas a computer achieved accuracies of 63.5%. However, the affect judges in that study were largely software developers. Perhaps higher classification accuracies could be obtained by humans trained in emotional intelligence, as in the case of clinical psychologists or FBI agents.

The next step in our research will be to combine the information from the different sensor channels into one emotion classifier that can be used in AutoTutor. We envision two possible methods to achieve this goal. The first option would be to acknowledge that each sensor is best at classifying a particular set of emotions. If so, the posture sensor would be responsible for detecting boredom and flow, facial feature tracking would be used for confusion and delight, and frustration would be classified on the basis of the dialogue features. The problem with this approach is that it does not leave any room for improvement because we are committed to the maximum accuracy values affiliated with each sensor.

Perhaps a more attractive alternative would be to combine features from different sensors to determine whether sensor fusion results in increased classification accuracy. One initial step at sensor fusion is to develop and test four additional classification models: dialogue + posture, dialogue + face, posture + face, and dialogue + posture + face. By repeating the analyses reported earlier, one could determine whether some emotions are best classified by considering features of each channel separately or by one of the four composite feature sets. We are currently testing the hypothesis that classification performance from multiple channels will exhibit *super-additivity*, i.e., performance superior to an additive combination of individual channels. An alternative hypothesis is that there will be *redundancy* across the channels, i.e., adding additional channels yields negligible incremental gains.

Once we have isolated the individual channel or combination of channels that maximizes the discriminability of each of the 5 affective states from neutral, we would require a super classifier to integrate the outputs of the individual affect-neutral classifiers. We envision a collection of affect-neutral classifiers that would first determine whether the incoming dialogue pattern resonated with any one or more of the emotions (versus a neutral state). If there is resonance with only one emotion, then that emotion would be declared as being experienced by the learner. If there is resonance with 2 or more emotions, then a conflict resolution module would be launched to decide between the alternatives. Perhaps this would be a second level affect classifier. We have indeed been encouraged by our preliminary experiments that calibrate the accuracy of such a multi-layered emotion classifier.

Classification of learner emotions is an essential step in building a tutoring system that is sensitive to the learner's emotions. The other essential component is to build mechanisms that empower AutoTutor to intelligently respond to these emotions, as well as to their states of cognition, motivation, social sensitivity, and so on. In essence, how can an affect-sensitive AutoTutor respond to the learner in a fashion that optimizes learning and engagement? At this point in the science, it is an open question as to what the optimal pedagogical strategies would be.

We have explored possible strategies that address the presence of boredom, frustration, flow, and confusion in the learner. If the learner is bored, a state that has been negatively correlated with learning[7], then AutoTutor should engage the learner in an activity that increases interest and cognitive arousal. These might include a simulation, options of choice, a challenge, or a seductive embedded game. According to the results presented in this study, frustration could be remedied with dialogue strategies that use more direct feedback, assertions, and corrections of misconceptions. According to the intuitions of many of our colleagues, an empathetic tutor would be effective in alleviating frustration. An efficient handling of the flow experience may be to lay low and optimally manage the flow. In this case, AutoTutor should continue its normal interaction and possibly provide new or more difficult content as old content is mastered. Confusion presents a key opportunity for the ITS to encourage learning. Since confusion has been positively correlated with learning[7], it is not in itself a state to avoid during the learning process. It might be best to allow the student to stay in a state of confusion for awhile. However, determining the appropriate level of confusion is not a straightforward computation and might ideally be adapted to the personality and amount of world knowledge of the learner. For example, academic risk theory contrasts (a) the adventuresome learners who want to be challenged with difficult tasks, to take risks of failure, and to manage negative emotions when they occur with (b) those learners who take fewer risks, avoid complex tasks, effectively minimizing learning situations in which they are likely to fail and experience negative emotions. Some of these variables can be easily measured by domain knowledge pretesting, academic record, and personality tests prior to the intervention.

The pedagogical strategies discussed above involve AutoTutor simply *reacting* to the emotions of the learner. However, this approach might not suffice if learners cycle through their emotions in a context-sensitive fashion. When learner's experience negative emotions of boredom and frustration, they are more likely to stay in these states rather than transition into the more positive states of flow and delight. In contrast, learners in a state of flow tend to remain engaged or alternatively transition into confusion, an affective state that is positively correlated with learning. Therefore, in order to optimize learning, AutoTutor may need to steer learners into a virtuous cycle of flow and confusion, while simultaneously avoiding the viscous cycle of boredom and frustration. This complex mechanism suggests that it may be important to move beyond the simple reactive strategy of detecting and responding to negative emotions. AutoTutor may also need to *proactively* anticipate and attempt to prevent the onset of these negative emotions that are detrimental to learning and engagement.

As we explore the relationship between learning and emotion, we anticipate that we will need to revise, redefine, and possibly reconceptualize our theoretical perspective and our learning environments. We are in the process of investigating various theoretical frameworks that relate learning and emotions. We are currently testing the simple framework with five affective experiences (boredom, confusion, delight, flow, and frustration), with an eye for testing the complex dynamic model proposed by Kort, Reilly, and Picard[2] and the psychological theories that link emotion and cognition proposed by Mandler, Stein, Piaget, Vygotsky, and others. However, the proof of the pudding is in the eating, i.e. how well do the theories account for the data collected in the multiple annotator study and its recent replication. Any revised, augmented, or future theories of learning and emotion can be tested by modifying AutoTutor with an appropriate set of probabilistic production rules. These rules systematically generate dialog moves that are differentially sensitive to affective states of the learner. The affect-sensitive AutoTutor provides a rigorous foundation for testing alternative scientific theories in addition to discovering which mechanisms end up producing the best learning gains and learner satisfaction.

Acknowledgements

We thank our research colleagues at the University of Memphis and MIT. We also acknowledge Steelcase Inc. for providing us with the Tekscan Body Pressure Measurement System at no cost. This research was supported by the National Science Foundation (REC 0106965 , ITR 0325428, and REC 0633918). Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

References

1. A. C. Graesser, P. Chipman, B. C. Haynes, and A. Olney, "AutoTutor: An intelligent tutoring system with mixed-initiative dialogue", *IEEE Transactions in Education*, 2005, vol 48, pp. 612-618.
2. B. Kort, R. Reilly and R. W. Picard, "An Affective Model of Interplay Between Emotions and Learning: Reengineering Educational Pedagogy-Building a Learning Companion", *Proc. 2nd IEEE Int'l Conf. Advanced Learning Technologies* (ICALT 2001), IEEE CS Press, 2001, pp. 0043.
3. C. Conati, "Probabilistic Assessment of User's Emotions in Educational Games", *Journal of Applied Artificial*

Intelligence, 2002, vol 16, no. 7-8, pp. 555-575.

4. D. J. Litman and S. Silliman, "ITSPOKE: An intelligent tutoring spoken dialogue system". Proc. 4th meeting of HLT/NAACL, 2004, pp. 52-54.
5. M. Pantic and L.J.M. Rothkrantz, "Towards an Affect-sensitive Multimodal Human-Computer Interaction", *IEEE Special Issue on Multimodal Human-Computer Interaction*, 2003, vol. 91, no. 9, pp. 1370-1390.
6. P. Ekman and W. V. Friesen, *The facial action coding system: A technique for the measurement of facial movement*, Consulting Psychologists Press, 1978.
7. S. D. Craig, A. C. Graesser, J. Sullins, and B. Gholson. "Affect and learning: An exploratory look into the role of affect in learning", *Journal of Educational Media,* 2004, vol. 29, pp. 241-250.
8. S. K. D'Mello, S. D. Craig, J. Sullins, and A. C. Graesser, "Predicting affective states through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue", *International Journal of Artificial Intelligence in Education,* 2006, vol. 16, pp. 3-28.
9. A. C. Graesser, B. McDaniel, P. Chipman, A. Witherspoon, S. D'Mello, and B. Gholson, "Detection of Emotions During Learning with AutoTutor", Proc. 28th Ann. Conf. Cognitive Science Soc, (COGSCI 2006), Cognitive Science Society, Inc., 2006, pp. 285-290.
10. R. el Kaliouby and P. Robinson, "Generalization of a vision-based computational model of mind-reading"*, 1st Intl. Conf. on Affective Computing and Intelligent Interaction,* LNCS 3784, Springer, 2005, pp 582-589.