Automatic Detection of Learner's Affect from Conversational Cues

SIDNEY K. D'MELLO[1], SCOTTY D. CRAIG[2],
AMY WITHERSPOON[3], BETHANY MCDANIEL[3], and ARTHUR GRAESSER[3]

[1] *Department of Computer Science, The University of Memphis*
Memphis, TN 38152, USA. Email: sdmello@memphis.edu
[2] *Learning Research and Development Center, University of Pittsburgh*
Pittsburgh, PA, 15260, USA. E-mail: scraig@pitt.edu
[3] *Department of Psychology, The University of Memphis*
Memphis, TN, 38152, USA., E-mail: {awthrpsn, btmcdanl, a-graesser}@memphis.edu

This submission is intended for the Special Issue on Affective Modeling and Adaptation. This paper (or a similar version) is not currently under review by a journal or conference, nor will it be submitted to such within the next three months.

Abstract

We explored the reliability of detecting a learner's affect from conversational features extracted from interactions with AutoTutor, an intelligent tutoring system that helps students learn by holding a conversation in natural language. Training data were collected in a learning session with AutoTutor, after which the affective states of the learner were rated by the learner, a peer, and two trained judges. Inter-rater reliability scores indicated that the classifications of the trained judges were more reliable than the novice judges. Seven data sets that temporally integrated the affective judgments with the dialogue features of each learner were constructed. The first four datasets corresponded to the judgments of the learner, a peer, and two trained judges, while the remaining three data sets combined judgments of two or more raters. Multiple regression analyses confirmed the hypothesis that dialogue features could significantly predict the affective states of boredom, confusion, flow, and frustration. Machine learning experiments indicated that standard classifiers were moderately successful in discriminating the affective states of boredom, confusion, flow, frustration, and neutral, yielding a peak accuracy of 42% with neutral (*chance*=20%) and 54% without neutral (*chance* = 25%). Individual detections of boredom, confusion, flow, and frustration, when contrasted with neutral affect, had maximum accuracies of 69%, 68%, 71%, and 78%, respectively (chance=50%). The classifiers that operated on the emotion judgments of the trained judges and combined models outperformed those based on judgments of the novices (i.e., the self and peer). Follow-up classification analyses that assessed the degree to which machine-generated affect labels correlated with affect judgments provided by humans revealed that human-machine agreement was on par with novice judges (self and peer) but quantitatively lower than trained judges. We discuss the prospects of extending AutoTutor into an affect-sensing intelligent tutoring system.

Key words: *Affect detection, human-computer interaction, human-computer dialogue, dialogue features, discourse markers, conversational cues, Intelligent Tutoring Systems, AutoTutor*

1. Introduction

Researchers in the field of human-computer interaction (HCI) have recently been modeling the affect of users (feelings, moods, emotions) in an attempt to develop more effective, user-friendly, and naturalistic applications (Bianchi-Berthouze & Lisetti, 2002; Hudlicka & McNeese, 2002; Klein, Moon, & Picard, 2002; Scheirer, Ferndandez, Klein, & Picard, 2002; Prendinger & Ishizuka, 2005; Whang, Lim, & Boucsein, 2003). The process of transforming a non-affect sensitive system into one that is responsive to its user's affective states involves the modeling of a cycle known as the *affective loop*. The affective loop is realized by the *detection* of the user's affective states, the s*election* of appropriate actions that encompass the user's affective states in the decision making process, and the *synthesis* of appropriate emotional expressions by the system. Given this functional specification of the affective loop, one of the fundamental challenges involves the robust detection of the user's affect.

There have been systematic attempts to automatically detect emotions by means of signal detection algorithms that operate on a host of sophisticated sensors. There has been some success in using physiological signals in emotion detection (Rani, Sarkar, & Smith, 2003; Picard, Vyzas, & Healey, 2001; Whang, Lim, & Boucsein, 2003), but a potential pitfall to this approach is the reliance on obtrusive sensing technologies, such as skin conductance, heart rate monitoring, and measurement of brain activities. While obtrusive detection of affect may be suitable in some domains after users habituate to the presence of these sensors, they are not satisfactory in environments in which the sensors distract users and interfere with the primary tasks. Therefore, we argue for the use of non-intrusive bodily sensors, such as cameras that track facial features and microphones that monitor acoustic-prosodic speech contours. The use of these sensors is not unique. A majority of the affect detection systems rely on facial feature tracking (Cohn & Kanade, in press; Oliver, Pentland, & Berand, 1997) and acoustic-prosodic vocal features (Bosch, 2003; Grimm et al., 2006; Litman & Forbes-Riley, 2004; Shafran & Mohri, 2005). To a lesser extent, some research has also focused on affect detection from posture patterns (Mota & Picard, 2003).

Our interest in the field of affect detection emerges from our desire to transform an intelligent tutoring system (ITS), AutoTutor, into an affect responsive system. AutoTutor is an intelligent tutoring system that helps learners construct explanations by interacting with them in natural language and helping them use simulation environments (Graesser, Chipman, Haynes, & Olney, 2005; Graesser, Person, Harter, & TRG, 2001). ITSs such as AutoTutor have implemented several systematic strategies for promoting learning, such as error identification and correction, building on prerequisites, frontier learning (expanding on what the learner already knows), student modeling (inferring what the student knows and having that information guide tutoring), and building coherent explanations (Aleven & Koedinger, 2002; Anderson, Corbett, Koedinger, & Pelletier, 1995; Gertner & VanLehn, 2000; Koedinger, Anderson, Hadley, & Mark, 1997; Lesgold, Lajoie, Bunzo, & Eggan, 1992; Sleeman & Brown, 1982; VanLehn, 1990) While ITSs have typically focused on the learner's cognitive states, we believe that they can be far more than mere cognitive machines. ITSs can be endowed with the ability to recognize, assess, and react to a learner's affective state. We are not alone in this view. For example, one of the first suggestions for endowing computer tutors with a degree of empathy or affect was made by Lepper and Chabay (1988). Consequently, we claim that intelligent tutoring systems should include a mechanism for motivating the learner, detecting the learner's emotional/motivational state, and appropriately responding to that state (Issroff & del Soldato, 1996; Lepper & Chabay, 1988; Lepper and Woolverton, 2002).

DeVincente and Pain (2002) have argued that motivation components are as important as cognitive components in tutoring strategies, and that important benefits would arise from considering techniques that track the learner's motivation and emotions. There is some evidence, for example, that tracking and responding to human emotions on a computer increases students' persistence (Aist et al., 2002). Kim (2005) conducted a study which demonstrated that the interest and self-efficacy of a learner significantly increased when the learner was accompanied by a pedagogical agent acting as a virtual learning companion that is sensitive to the learner's affect. Linnenbrink and Pintrich (2002) reported that the posttest scores of physics understanding decreased as a function of negative affect during learning.

An emotionally-sensitive learning environment, whether it be human or computer, requires some degree of accuracy in classifying the learner's affective states. The emotion classifier need not be perfect but must have some modicum of accuracy. While the larger project of integrating affect-sensing capabilities into AutoTutor makes use of facial feature tracking, speech analyses, and posture patterns for affect detection (D'Mello, Craig, Gholson, Franklin, Picard, & Graesser, 2005), this paper focuses on detecting affect from discourse features obtained from AutoTutor's natural language mixed-initiative dialogue. Although dialogue has traditionally been a relatively unexplored channel for affect detection, it is a reasonable information source to explore because dialogue information is abundant in virtually all conversations and is inexpensive to collect.

A growing body of research has investigated emotions in human-human dialogues (Alm & Sproat, 2005; Forbes-Riley & Litman, 2004) and human-computer dialogues (Litman & Forbes-Riley, 2004), although the literature on automated affect detection from the latter is somewhat sparse. Dialogue (i.e., discourse) features have typically been used in conjunction with acoustic-prosodic and lexical features obtained through an interaction with spoken dialogue systems. The lexical features usually are restricted to either human or automatic transcriptions (with speech recognition engines) of user utterances. The acoustic-prosodic features are composed of vocal cues that typically include speech rate, intonation, and volume. A number of research groups have reported that appending an acoustic-prosodic and lexical feature vector with dialogue features results in a 1-4% improvement in classification accuracy (Ang, Dhillon, Krupski, Shriberg, & Stolcke, 2002; Litman & Forbes-Riley, 2004; Liscombe, Riccardi, & Hakkani-Tür, 2005; Lee and Narayanan, 2004). A classic example involving this use of dialogue features is work investigating dialogue and emotions conducted on the program ITSPOKE (Litman, Rose, Forbes-Riley, VanLehn, Bhemhe, Silliman, 2004; Litman & Silliman, 2004). ITSPOKE integrates a spoken language component into the Why2-Atlas tutoring system (VanLehn et al., 2002). With ITSPOKE, Litman and Forbes-Riley (2004) analyzed spoken student dialogue turns on the basis of lexical and acoustic features, with codings of negative, neutral or positive affect. They were able to reach high levels of accuracy in detecting affect categories.

In a similar vein, Ang et al. (2002) reported that the inclusion of discourse features, such as the current turn within a session and the associated dialogue acts of the current turn, resulted in a 4% improvement in performance over lexical and prosodic features. Their research involved detecting annoyance and frustration within the context of a travel reservation system. Similarly, Liscombe, Riccardi, and Hakkani-Tür (2005) reported that the use of dialogue features caused a 1.2% improvement over the use of acoustic-prosodic and lexical features in discriminating between positive and negative emotions. Their results were obtained by analyzing a large database of 5,690 spoken utterances obtained from user interactions with the *How May I Help You* spoken dialogue system (Gorin, Riccardi, & Wright, 1997). An additional 2.8% improvement in accuracy was obtained by the inclusion of contextual features spanning two previous turns. Another important example of the use of dialogue for affect detection is provided by Lee and Narayanan (2004). They reported that the use of dialogue features of user utterances obtained from a call center produced a 3% increase in accuracy over prosodic and lexical features in discriminating between negative and non-negative emotions.

Innovative uses of dialogue have emerged from research on the identification of problematic points in human-computer interactions (Batliner, Fischer, Huber, Spilker, and Noth, 2003; Carberry, Lambert, and Schroeder, 2002; Walker, Langkilde-Geary, Hastie, Wright, & Gorin, 2002). For example, Carberry, Lambert, and Schroeder (2002) proposed an algorithm to recognize doubt by examining linguistic and contextual features of dialogue in conjunction with world knowledge. Batliner et al. (2003) reported that discourse information resulted in a 1.2% improvement in classification accuracy over lexical and prosodic features alone.

We can identify three major differences between our approach to affect detection reported in this paper and some of the earlier research involving the use of dialogue to detect affect. The first difference is that we explore a larger array of discourse variables. We believe dialogue can be a serious competitor to more popular measures of user affect, such as facial and acoustic-prosodic features. The second difference is that previous efforts investigating dialogue were limited to a small set of affective states, such as neutral, negative, and positive (Litman & Forbes-Riley, 2004), negative versus positive/non-negative (Liscombe, Riccardi, & Hakkani-Tür, 2005; Lee and Narayanan, 2004), or annoyance versus frustration (Ang et al., 2002). These contrasts may be suitable for some domains, but they are not sufficient to encompass a realistic gamut of learning (Conati, 2002). Additional complexities arise from the fact that a person's reaction to the presented material can change as a function of their goals, preferences, expectations and knowledge state. Consequently, our research involves the detection of a larger set of affective states within the arena of complex-learning. The relevant emotions (i.e., affective states) include boredom, confusion, delight, flow, frustration, neutral, and surprise. The third difference between this research and other efforts is the method of establishing ground-truth categories of affect. A number of researchers have relied on a single operational measure when inferring a learner's emotion, such as self reports (De Vicente & Pain, 2002; Klein, Moon, & Picard, 2002; Matsubara & Nagamachi, 1996) or ratings by independent judges (Liscombe, Riccardi, & Hakkani-Tür, 2005; Litman & Forbes-Riley, 2004; Mota & Picard, 2003). In contrast, we propose the combination of several different measures of a learner's affect. Our measures of emotion incorporate judgments made by the learner, a peer, and two trained judges, as will be elaborated later.

We begin by describing the AutoTutor learning environment and exploring the interplay between emotions and learning. In particular, we describe three studies, two of which were used to isolate the set of affective states that accompany complex learning. The third study tackles the problem of human measurement of emotions. The data has served as training and testing data for a number of machine learning algorithms, with the affect ratings representing the gold standard. The subsequent section describes the AutoTutor dialogue features and provides a synopsis of some of our past research efforts in detecting the learner's emotions from these features. The Results section begins with a series of statistical analyses that evaluate the hypothesis that dialogue features can significantly predict the learner's affect. Two dimensionality reduction techniques are subsequently investigated as preprocessing techniques for the machine learning algorithms. The machine learning experiments attempt to assess the reliability of automatically detecting the learner's affect from AutoTutor's dialogue. We conclude by discussing the prospect of integrating additional sensors (cameras, posture sensors, etc.) in an effort to boost classification accuracy. Our ultimate goal is to explore how the learner's affective states may be integrated into AutoTutor's pedagogical strategies and thereby improve learning.

2. The Relationship between Affect and Complex Learning

Empirical research on emotions is quite extensive, but the scientific literature on the relation between emotions and cognition is considerably sparse and scattered, especially in relation to complex learning (Goleman, 1995; Mandler, 1984; Stein & Levine, 1991). In the popular science literature, Goleman's (1995) book *Emotional Intelligence* raises the question of how emotions impact learning, and vice versa. However, the book did not address many of the important research questions to guide future empirical work. Our previous research has identified the affective states that accompany learning introductory computer literacy (Craig et al., 2004; D'Mello et al., 2006) and some alternative ways to measure these affective states (Graesser et al., 2006). These previous efforts are described in this section. The desire to elicit affect from a non-traditional source like conversational dialogue raises additional issues that we also address in this section.

2.1. The AutoTutor Learning Environment

The Tutoring Research Group (TRG) at the University of Memphis developed AutoTutor, a fully automated computer tutor that simulates human tutors and holds conversations with students in natural language (Graesser et al.; 2001; 2005a; 2005b). AutoTutor attempts to comprehend the students' natural language contributions and then responds to the students' typed input with adaptive dialogue moves similar to human tutors. AutoTutor helps students learn by presenting challenging problems (or questions) from a curriculum script and engaging in a mixed-initiative dialogue while the learner constructs an answer.

AutoTutor has different classes of dialogue moves that manage the interaction systematically. AutoTutor provides *feedback* on what the student types in (positive, neutral, or 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 *misconceptions* and erroneous ideas, *answers* the student's questions, and *summarizes* topics. A full answer to a question is eventually constructed during this dialogue, which normally takes between 30 and 100 turns between the student and tutor for one particular problem or main question.

As an example, the exchange below is one of the shorter tutorial dialogues between a college student and AutoTutor. This student was an above-average student in the computer literacy class.

1. AUTOTUTOR: When you turn on the computer, how is the operating system first activated and loaded into RAM? [this is the *main question*]
2. STUDENT: When the computer is turned on a file is automatically booted up.
3. AUTOTUTOR: Anything else? [this is a *pump* for more information]
4. STUDENT: This file contains the operating system.
5. AUTOTUTOR: What about storage? [this is a *hint*]
6. STUDENT: The operating system is then read into RAM and activated.
7. AUTOTUTOR: Right. The CPU executes instructions in ROM, which helps the operating system to start. [after giving *positive feedback*, there is an *assertion*]
8. STUDENT: After being read from ROM the OS is activated.
9. AUTOTUTOR: Right. Let's review what we've just gone over. The operating system must first be stored on the hard disk. When you turn on the computer, the CPU executes instructions stored in ROM. These instructions help boot the operating system. The operating system is then loaded from the hard disk into RAM. [after *positive feedback* comes a *summary*]

AutoTutor's knowledge about the topic being tutored (computer literacy in this study) is represented by Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) and a moderately structured curriculum script on the material. LSA is a statistical technique that measures the conceptual similarity of two texts on the basis of a higher-dimensions space with a few hundred dimensions (Landauer & Dumais, 1997). The dimensions are statistically derived from a large corpus of texts and the words that appear in the documents of the corpus. LSA computes a geometric cosine (typically ranging from 0 to 1) that represents the conceptual similarity between the two texts. In AutoTutor, LSA is used to assess the quality of student responses (by comparing student input to expected content) and to monitor other informative parameters, such as topic coverage and student ability level.

AutoTutor has been tested in several experiments on over 2500 students in computer literacy and physics courses. Significant learning gains were obtained in all of these experiments (an average sigma of .8), particularly at the level of deep explanations as opposed to shallow facts and ideas (VanLehn, Graesser, Jackson, Jordan, Olney, & Rose, in press). AutoTutor has also been evaluated on the conversational smoothness and the pedagogical quality of its dialogue moves in the turn-by-turn tutorial dialogue (Person, Graesser, & TRG, 2002). Person, Graesser and TRG (2002) performed a *bystander Turing test* on the naturalness of AutoTutor's dialogue moves. Bystanders were unable to discriminate between dialogue moves of AutoTutor and dialogue moves of a real human tutor.

2.2. Identifying the Affective States that Accompany Complex Learning

There have been theories that link cognition and affect very generally, such as those of Mandler (1984), Bower (1981), Stein and Levine (1991), Ortony, Clore, and Collins (1988), and more recently by Russell (2003). While these theories convey general links between cognition and emotions, they do not directly explain and predict the sort of emotions that occur during complex learning, such as attempts to master physics, biology, or computer literacy. Some emotions presumably have a more salient role in learning than others (Linnenbrink & Pintrich, 2004). What are these emotions? How are they linked to cognition? These are the fundamental questions that surface in modeling affect in learning environments.

Ekman and Friesen (1978) have proposed 6 *basic* emotions that are ubiquitous in everyday experience. The six basic emotions include fear, anger, happiness, sadness, disgust, and surprise. However, many have called into question the relevance of these basic emotions to the learning process (Kort, Reilly, & Picard, 2001). Some researchers have argued for a different set of emotions that influence learning and cognition, namely boredom (Csikszentmihalyi, 1990; Miserandino, 1996), confusion (Graesser & Olde, 2003; Kort, Reilly, & Picard, 2001), delight (Fredrickson & Branigan, 2005; Silvia & Abele, 2002), flow (Csikszentmihalyi, 1990), frustration (Kort, Reilly, & Picard, 2001; Patrick et al, 1993), and surprise (Schutzwohl & Borgstedt, 2005). In an earlier study, we reported that increased levels of boredom were negatively correlated with the learning of computer literacy, whereas increased levels of confusion and the state of flow (being absorbed in the learning process, Csikszentmihalyi, 1990) were positively correlated with learning in an AutoTutor learning environment (Craig, Graesser, Sullins, & Gholson, 2004).

In a recently completed study, we adopted an emote-aloud procedure (D'Mello, Craig, Sullins, & Graesser, 2006), a variant of the think-aloud procedure (Ericsson & Simon, 1993), as an online measure of the learners' affective states during learning. College students were asked to state the affective states they were feeling while working on a task, in this case being tutored in computer literacy with AutoTutor. This method allowed for on-line identification of emotions while working on a task with minimal task interference. Seven participants were run in the emote-aloud study and the procedure yielded 215 emote-alouds. The emotions of interest in the emote-aloud study were anger, boredom, confusion, eureka, frustration, contempt, curious, and disgust, but only boredom, confusion, eureka, and frustration were frequently reported by the participants. Additionally, although eureka was reported with a modest frequency, we suspect that this response may have functionally signified delight from giving a correct answer or surprise from getting unexpected positive feedback from the tutor rather than a deep eureka experience (i.e., a flash of insight, followed by extremely positive affect). In light of these findings, we have refined our list of emotions affiliated with deep learning to boredom, confusion, flow, and frustration. Additionally, three new affective states (neutral, delight, and surprise) have been added to the list. Neutral was added because unlike the emote-aloud study in which participants voluntarily expressed their affective states, the current study forced participants to state their emotions every 20 seconds (described below). Delight and surprise were added as functional replacements for eureka.

2.3. Human-Measurement of Emotions: The Multiple Annotator Study

Modeling affect involves determining what emotion a learner is experiencing at particular points in time. Emotion is a construct (i.e., an inferred conceptual entity), so one can only approximate its true value. Researchers sometimes have relied on a single operational measure in inferring a learner's emotion, such as self reports (De Vicente & Pain, 2002; Klein, Moon, & Picard, 2002; Matsubara & Nagamachi, 1996) or ratings by independent judges (Liscombe, Riccardi, & Hakkani-Tür, 2005; Litman & Forbes-Riley, 2004; Mota & Picard, 2003), but we propose the combination of several different measures of a learner's affect. Our measures consist of emotion judgments made by the learner, a peer, and two trained judges. Employing multiple measures of affect is compatible with the standard criterion for establishing convergent validity (Campbell & Fiske, 1959).

We conducted a study which consisted of 28 participants interacting with AutoTutor for 32 minutes on one of three randomly assigned topics in computer literacy: 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 (Morimoto, Koons, Amir, & Flickner, 1998). Posture patterns were captured by the Tekscan® Body Pressure Measurement System (Tekscan, 1997). A screen-capturing software program called Camtasia Studio (developed by TechSmith) was used to capture the audio and video of the participant's entire tutoring session with AutoTutor. The captured audio included the speech generated by the AutoTutor agent.

The affect judging process was conducted by synchronizing and displaying to the judges the video streams from the screen and the face. 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). Additionally, if participants were experiencing more than one affective state in a 20-second block, judges were instructed to mark each state and indicate which was most pronounced. However, in these situations only the more prominent affective state was considered in the current analyses. At the end of the study participants were asked to identify any affective states they may have experienced that were not included in the specified list of 7 emotions. However, a cursory look at this data did not reveal any new affective states.

Four sets of emotion judgments were made for the observed affective states of each participant's AutoTutor session. First, for the *self* judgments, the participant watched his or her own session with AutoTutor immediately after having interacted with the tutor. Second, 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 additional judges (called *trained judges*) judged all of the sessions individually; these trained judges had been trained on how to detect facial action units according to Paul Ekman's Facial Action Coding System (FACS) (Ekman & Friesen, 1978). The trained judges also had considerable experience interacting with AutoTutor. Hence, their emotion judgments were based on contextual dialogue information as well as the FACS system.

A list of the affective states and definitions was provided for all judges. The states were frustration, confusion, flow, delight, surprise, boredom, and neutral. Frustration was defined as dissatisfaction or annoyance. Confusion was defined as a noticeable lack of understanding, whereas flow was a state of interest that results from involvement in an activity. Delight was a high degree of satisfaction. Surprise was defined as wonder or amazement, especially from the unexpected. Boredom was defined as being weary or restless through lack of interest. Neutral was defined as no apparent emotion or feeling.

We examined 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 since the participants in this study were typically low domain knowledge students as indicated by their low pretest scores.

Interjudge reliability was computed using Cohen's kappa for all possible pairs of judges: self, peer, trained judge1, and trained judge2. Cohen's kappa measures the proportion of agreements between two judges with correction for baserate levels and random guessing. There were 6 possible pairs altogether. The kappas were reported in Graesser et al. (2006): self-peer (.08), self-judge1 (.14), self-judge2 (.16), peer-judge1 (.14), peer-judge2 (.18), and judge1-judge2 (.36). While these kappas appear to be low, they are on par with data reported by other researchers who have assessed identification of emotions by humans (Ang et al., 2002; Grimm et. al., 2006; Litman & Forbes-Riley, 2004; Shafran, Riley, & Mohri, 2003).An ANOVA performed on these 6 interjudge kappa revealed that there were significant differences in the inter-judge reliability scores among the six pairs, $F(5, 135) = 33.34$, $MSe = .008$, $p < .01$. Post hoc tests revealed that the self-peer pair had the lowest inter-judge reliability 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 peers are not particularly good at detecting learner emotions. Another conclusion is that training on Ekman's facial action coding system can enhance the reliability and accuracy of judgments of affective states.

Interrater reliability scores for individual emotions between the two trained judges revealed that delight and confusion has the highest kappas (.71 and .40 respectively). The kappa scores for flow (.30), neutral (.30), surprise (.27), and boredom (.26) were very similar and quantitatively higher than frustration (.21) which had the lowest kappa. On the basis of these results, we can infer that the trained judges had less difficulty in detecting emotions that are typically embodied with animated facial expressions such as delight and confusion. However, our data does suggest that frustration and surprise seem to be exceptions to this rule. The low kappa scores associated with frustration could perhaps be explained by the fact that most learners attempt to disguise frustration since it is considered to be a negative affective state. Moderate kappa scores for surprise could be explained by the low frequency of occurrence of this emotion. In fact trained judges only observed surprise in half of the participants and obtained a non-zero kappa score in about a third of these participants.

3. Synopsis of Prior Research on Affect and Dialogue

3.1. Features of AutoTutor's Mixed-Initiative Dialogue

A session with AutoTutor is comprised of a set of subtopics (main questions) 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 (e.g., hints, prompts, assertions), misconceptions, corrections of misconceptions, and other slots in the curriculum script that need not be addressed here. The expectations are ideally covered by a series of turns in AutoTutor's conversation with the student in an attempt to help the student construct an answer to the current main question (subtopic). When an acceptable answer with the appropriate details is gleaned from the student's responses (usually after 30 – 100 turns), 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, a variety of assessments of the response, the feedback provided, and the tutor's next move. Temporal information, such as the student's reaction time and response time, is also recorded. Table I provides an overview of relevant information channels that are available in AutoTutor's log files of the interaction history.

3.1.1 Temporal Information. The temporal information can be viewed as a combination of global and local temporal markers that span the period of interaction. The real time measures the time of a dialogue event in the tutoring session, and is measured in milliseconds but rounded to seconds for ease of interpretation. The subtopic number indicates the number of main questions answered. It provides a global measure of sequential position within the entire tutorial session. For example, for a one-hour session covering three subtopics, the third subtopic would indicate that the student is approximately in the 40-60 minute time span. The turn number, on the other hand, provides a local temporal measure. It is the nth turn of the student in the current question (subtopic). Finally, the student response time is the elapsed time (in milliseconds converted to seconds for easy interpretation) between the verbal presentation of the question by AutoTutor and the student submitting an answer.

3.1.2 Response Information. AutoTutor uses LSA for the majority of its assessments of the student's responses to a question, as will be discussed below. Another measure we consider is the *verbosity* of the student's responses. The verbosity is measured by the *number of words* and the *number of characters* in the student's response. A recent measure of the student's response to AutoTutor is based on a classification of the student's response (SAC) according to a Speech Act Classification system (Olney, Louwerse, Mathews, Marineau, Hite-Mitchell, & Graesser, 2003). The system classifies each response into one of a number of categories; those of interest in this research involve topic-unrelated *frozen expressions* (e.g., I don't know, What did you say?, coded as -1) and topic-related *contributions* (scored as a 1).

3.1.3 Answer Quality Assessments. AutoTutor relies on LSA (Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, in press) as its primary computation of the quality of student responses in student turns. The local assessments for a given turn N measures the student's response for that turn on the basis of its similarity to good answers (expectations) and bad answers (misconceptions and bugs). The *local good score* is the highest match score between the content of student turn N and the set of expectations representing good answers. The local *bad score* is the highest match to the set of bad answers. A high local good score reflects progress in answering the main question, whereas a high local bad score reflects resonance with misconceptions. The *delta local good score* and the *delta local bad score* measure changes in the local good score and the local bad score, respectively, compared with student turn N-1.

The four Global parameters (see Table I) perform the same assessments as the local parameters with the exception that the text used for the LSA match is an aggregation of all of the student's turns (1 through N) for a given subtopic. With this scheme, a student's past responses to a subtopic are considered in AutoTutor's assessment of the student's current response.

Table I. Description of the information mined from AutoTutor's log files at the end of each student turn

Channel	Sub channel	Description
Temporal Information	Real Time	Time in seconds since the beginning of the session
	Subtopic No.	The current subtopic (question) in this session
	Turn No.	The no. of the conversation turn within a subtopic
	Response Time	Time between question and answer submission
Response Information	No. of words	The number of words in the student's response
	No. of chars	The number of characters in the student's response
	Speech Act	Speech Act category of the student's response
Answer Quality Assessment	Local Good	Similarity of student's response to an expectation
	Delta Local Good	The change in the Local Good Score
	Global Good	Similarity of response history to expectations
	Delta Global Good	The change in the Global Good Score
	Local Bad	Similarity of student's response to a bad answer
	Delta Local Bad	The change in the Local Bad Score
	Global Bad	Similarity of response history to bad answers
	Delta Global Bad	The change in the Global Bad Score
Tutor Directness	Pump	Minimal information provided. e.g. "What else"
	Hint	Provides a hint to the student to fill in proposition
	Prompt	Prompts student to fill in a missing content word
	Correction	Corrects the student's misconception
	Assertion	Asserts information about an expectation
	Summary	Provides a summary of the answer
Tutor Feedback	Positive	Provides feedback terms such as: "good job",
	Neutral Positive	Provides feedback terms such as: "yeah", "right"
	Neutral	Provides feedback terms such as: "uh huh",
	Neutral Negative	Provides feedback terms such as: "kind of"
	Negative	Provides feedback terms such as: "wrong", "no"

Note. The various sub channels for tutor directness and tutor feedback channels are ordered onto two individual scales. Therefore, the number of dialogue predictors is taken to be 17 and not 26.

3.1.4 Tutor Directness. At the end of each student turn, AutoTutor incorporates the various LSA assessments when choosing its next pedagogically appropriate dialogue move. When AutoTutor tries to get a single expectation (*E*) covered (e.g., "The hard disc is a storage medium"), this goal is posted and is achieved by AutoTutor presenting a series of different dialogue moves across turns until the expectation *E* is expressed by the student or as a last resort by the tutor. It first gives a *pump* (What else?), then a *hint* (What about the hard disk?), then a *prompt* for a specific important word (The hard disk is a medium of what?), and then simply *asserts* the information (The hard disc is a medium for storage). After all of the expectations for the problem are covered, a *summary* is provided by AutoTutor. Given this mechanism of encouraging the student to cover the expectations, the dialogue moves chosen can be ordered on a *directness* scale (ranging from -1 to 1) on the basis of the amount of information AutoTutor supplies to the learner. The ordering is *pump* < *hint* < *prompt* < *assertion* < *summary.* A pump conveys the minimum amount of information (on the part of AutoTutor) whereas a summary conveys the most amount of explicit information.

3.1.5 Tutor Feedback. AutoTutor's short feedback (positive, neutral, negative) is manifested in its verbal content, intonation, and a host of other non-verbal conversational cues. Table I shows examples of AutoTutor's responses, characterized by the type of feedback being provided. Similar to the directness scale constructed above, AutoTutor's feedback was mapped onto a scale ranging from -1 (negative feedback) to 1 (positive feedback).

3.2. Relating Affect and Dialogue

The first investigation into the potential of the dialogue features being a viable channel for affect detection was performed on the data from the emote-aloud study (D'Mello et al., 2006). In that study, college students verbally expressed their emotions during interactions with AutoTutor. The AutoTutor log files were mined to obtain information from the various dialogue channels described above. Each dialogue feature vector was then associated with an emotion category on the basis of the verbalized affect ratings. Correlation and regression analyses confirmed the hypothesis that dialogue features could significantly predict the affective states of confusion, eureka (delight), and frustration but not boredom. Standard classification techniques used to assess the reliability in discriminating between confusion, eureka, and frustration from the conversation features yielded an average accuracy of 58% (chance= 34.5%, D'Mello et al., 2006).

However, the D'Mello et al. (2006) results should be interpreted with caution because of a number of shortcomings identified with the emote-aloud procedure and the accompanying data analyses. The first limitation with the procedure was that there were floor effects in the reporting of some affective states. Consequently, we ended up concentrating on only the affective states of boredom, confusion, eureka/delight, and frustration. A second limitation was that a statistically significant model was not discovered for the affective state of boredom. The third limitation was that significant relationships between AutoTutor dialogue and the affective states were discovered only with three out of the five main channels of dialogue information available (see next section). The number of participants utilized in the emote-aloud study (N=7) was also a matter of some concern.

In light of the above discussion, we conclude that it is reasonable to explore the detection of learners' affect from the dialogue features. However, in order to consider dialogue as a serious competitor to the more popular bodily measures, such as the tracking of facial features and acoustic-prosodic features from speech, a more comprehensive evaluation of affect detection reliabilities from dialogue are required.

4. Results of Present Study

The present study evaluated whether conversational dialogue features are a viable channel for affect detection. The data used for the analyses was from the multiple annotator study in which 28 participants interacted with AutoTutor on topics in computer literacy. When aggregated across each 32-38 minute session for each of the 28 participants, we obtained 1470 student-tutor interaction turns and 2967, 3012, 3816, and 3723 emotion judgments for the self, peer, trained judge 1, and trained judge 2, respectively. A dialogue feature vector based on the features listed in Table I was then extracted for each student-tutor interaction turn. The feature vector was then associated with an emotion category on the basis of the human judges' affect ratings. More specifically, the emotion judgment that immediately followed a dialogue move (within a 15 second interval) was bound to that dialogue move. This data collection procedure yielded four ground truth models of the learner's affect, so we were able to construct 4 labeled data sets.

Affect judgment reliabilities between the human judges presented above revealed that the highest agreement was obtained between the trained judges (kappa = .36). However, it is still not firmly established whether the trained judges or the self judgments are closer to the ground truth. We address this issue by combining affect judgments from the four judges in order to obtain a better approximation of the learner's emotion. In particular, one data set was constructed on the basis of judgments in which both trained judges agreed. Another was constructed for judgments in which any two (or more) judges agreed. Similarly, a third data set was constructed for the affect judgments in which three (or more) judges agreed. A fourth additional data set was constructed for judgments in which all four judges agreed, but were eliminated from the subsequent analyses because of a very small sample size (N = 66). The frequencies of the emotions in each data set are listed in Table II.

Table II. Frequency of affective states in each data set

Affect Judge	**Frequency of Affective States**							**Sum**
	Boredom	*Confusion*	*Delight*	*Flow*	*Frustration*	*Neutral*	*Surprise*	
Self	164	172	52	176	167	265	28	1024
Peer	189	172	25	177	97	344	36	1040
Trained Judge 1	104	258	82	190	139	321	21	1115
Trained Judge 2	242	238	67	102	58	390	12	1109
Trained Judges	81	150	61	67	62	196	6	623
Two Agree	144	167	42	105	84	326	6	874
Three Agree	64	90	22	49	30	154	1	410

4.1 Multiple Regression Analyses

Multiple regression analyses were conducted to determine the extent to which the seven affective states of interest could be predicted from the various dialogue features. For each of the seven data sets (self, peer, trained judge1, trained judge2, trained judges agree, any 2 judges agree, and any 3 judges agree), seven multiple regression models were constructed, one for each of the affective states, yielding 49 models in all. The criterion variable for each multiple regression analysis was the affective state (1 or 0 if present or absent respectively) whereas the predictor variables were the set of dialogue features. It is widely acknowledged that strongly correlated predictor variables tend to cause instability in multiple regression models. Therefore, we first identified and subsequently eliminated the collinear dialogue features. We adopted a correlation threshold (Pearson's $r > .7$) to remove collinear predictor variables. Specifically, when two variables were identified as collinear predictors, the one with a stronger overall correlation with the affective states was preserved. This approach reduced the 17 features to 11 features by discarding real time, local bad score, global bad score, delta local bad score, delta global bad score, and number of words.

In order to partial out variability among participants, the multiple regression analyses were conducted in two steps. In step 1, the predictors included the participants' pretest scores and dummy coded variables to differentiate participants. In step 2, the group of predictors was the 11 different conversation features after six potential predictors were excluded as a result of the collinearity analysis. Step 1 was entered first, with the residual variance passed onto step 2. In this fashion, we could partial out any of the variability associated with the participants' characteristics and determine the unique variance that could be ascribed to particular conversation features.

4.1.1 Analysis of Regression Models. Statistically significant overall relationships (at the $p < 0.05$ level) were discovered for boredom, confusion, flow, frustration, and neutral, but not for delight and surprise. In the cases of delight and surprise, step 1 was usually significant but not step 2. Such a result implies that the various conversation features were unable to add a significant improvement in classifying affect states above and beyond participant characteristics.

Table III. Summaries of the multiple regression models for emotions in each data set

Rating Type	**Model**	**df1,df2**	**Affective States**									
			Boredom		*Confusion*		*Flow*		*Frustration*		*Neutral*	
			R^2_{adj}	F	R^2_{adj}	F	R^2_{adj}	F	R^2_{adj}	F	R^2_{adj}	F
Self	PC	27,996	.134	6.86	.123	6.33	-	-	.129	6.59	.298	17.12
	PC+DF	11,985	.162	4.00	.171	6.26	-	-	.161	4.46	.315	3.18
Peer	PC	27,1012	.162	8.44	.097	5.14	-	-	.085	4.57	.275	15.63
	PC+DF	11,1001	.208	6.39	.107	1.98	-	-	.116	4.26	.287	2.44
Trained Judge1	PC	27,1087	.072	4.18	.032	2.37	.098	5.50	.025	2.04	.013	1.54
	PC+DF	11,1076	.122	6.70	.082	6.33	.194	12.80	.107	10.15	.029	2.67
Trained Judge2	PC	27,1081	.048	3.06	.036	2.51	.046	2.98	.054	3.34	.075	4.31
	PC+DF	11,1070	.128	10.03	.140	12.91	.135	11.09	.103	6.38	.105	4.32
Trained Judges	PC	27,595	.062	2.53	.040	1.96	.082	3.05	.039	1.95	.092	3.34
	PC+DF	11,584	.159	7.22	.123	6.13	.178	7.37	.138	7.19	.117	2.51
Two Judges	PC	27,846	.082	3.88	.064	3.21	.080	3.83	.063	3.19	.057	2.97
	PC+DF	11,835	.160	8.17	.159	9.67	.147	7.03	.138	7.70	.097	4.36
Three Judges	PC	27,382	.128	3.22	.114	2.96	.084	2.39	.120	3.07	.161	3.92
	PC+DF	11,371	.245	6.40	.229	6.16	.196	5.85	.164	2.84	.186	2.04
Mean R^2_{adj}	PC		0.098		0.072		0.078		0.074		0.139	
	DF		0.071		0.072		0.092		0.059		0.024	
	PC+DF		0.169		0.144		0.170		0.132		0.162	

PC: Regression model with participant characteristics only.
PC + DF: Regression model with participant characteristics and dialogue features.
All models statistically significant at the $p < .05$ level.

A number of conclusions can be drawn from the characteristics of the regression models presented in Table III. For the affective state of boredom, when aggregated across all 7 models, our features explained about 16.9% of the predictable variance, with 7.1% of the variance being accounted for by the step 2 conversation features alone (see last row of Table III). Similarly, on average 14.4% of the variance of affect classification was explained for confusion, with 7.2% obtained from the conversation features. For flow and frustration, the conversational features accounted for 9.2% and 5.9% of the total variances of 17% and 13.2% respectively. Additionally, the data sets based on face value judgments of the self and peer failed to converge on a statistically significant model for flow. The flow emotion is difficult to detect from the affect ratings made by the novice judges. Finally, the weakest model was obtained for neutral with the dialogue feature, explaining only 2.4% of the total variance of 16.2%. This implies that the dialogue features may not be very successful in discriminating the other affective states from neutral.

4.1.2 The Relationship between Dialogue Features and Learner's Affect. We can glean a number of generalizations regarding the relationship between dialogue and affective states, based on numerical direction (i.e. signs, +-) of the statistically significant coefficients of the multiple regression models (see Table IV). A number of relationships surfaced when one considers the significant predictors of the affective states where at least two judges agreed. In particular, boredom occurs later in the session (high subtopic number), after multiple attempts to answer the main question (high turn number), and when AutoTutor gives more direct dialogue moves (high directness). Alternatively, confusion occurs earlier in the session (low subtopic number), within the first few attempts to answer a question (low turn number), with slower responses (long response time), shorter responses (less characters), low quality answers (low local good LSA scores), with frozen expressions (negatively coded speech acts), when the tutor is less direct in providing information, and when the tutor provides negative feedback. The analyses indicated that flow occurs within the first few attempts to answer a question (low turn number), with quicker, longer, proficient responses (low response time, more characters, and high local good LSA score respectively), and is accompanied by positive feedback from the tutor. Frustration was prevalent later in the temporal span of a session (high subtopic number), with longer response times, with good answers towards the immediate question (high local good score), but poor answers towards the broader topic (low global good score), and negative tutor feedback.

Table IV. Significant predictors for the multiple regression models for emotions in each data set

Dialogue Features	**Affective States**																																		
	Boredom							*Confusion*							*Flow*							*Frustration*							*Neutral*						
	S	P	J1	J2	JA	2	3	S	P	J1	J2	JA	2	3	S	P	J1	J2	JA	2	3	S	P	J1	J2	JA	2	3	S	P	J1	J2	JA	2	3
Subtopic No.	+	+	+	+	+	+	+	-	-	-	-	-	-	-		-						+		+						-					
Turn No.	+	+	+	+	+	+	+		-				-	-				-		-			+												
Response Time											+	+	+	+				-	-	-	-					+		+			-		-	-	
No. Characters								-			-	-	-	-	+		+	+	+	+	+														
Global Good																							-	-	-	-									
Del. Glbl. Good																																			
Local Good			-								-	-	-				+			+				+	+	+	+								
Del. Local Good				-																															
Speech Act	-							-	-	-	-	-	-	-															+			+		+	+
Directness			+	+	+					-	-																								
Feedback					+					-			-				+	+	+	+		-	-	-	-	-	-	-	+	+	+	+	+	+	

S: Self Judgments, P: Peer Judgments, J1: Trained Judge1, J2: Trained Judge2
JA: Both trained judges agree, 2: Any two judges agree, 3: Any three judges agree
+ or - indicates that the feature is a positive or negative predictor in the multiple regression model at $p < .05$ significance level.
Empty cells indicate that a feature was not a statistically significant predictor for the respective emotion.

The relationships between the various dialogue features and the affective states described above are generally intuitive and in the expected directions. Of particular interest are the features that predict the affective state of frustration. In the case of frustration, the learner tends to have not been doing well on the topic in general, as indicated by a negative global LSA score, but they have taken a longer time to answer the question (increased response time) and have given a good answer to the immediate question (higher LSA local good score). However, AutoTutor's internal model of the learner's interaction has erroneously classified the student as being a poor learner and responds with increased negative feedback, which in turn increases frustration. It should be noted that in this case it would appear that the learner generally made an effort, as indicated by the increased response time and higher local good score, but did not receive the positive response expected. This would suggest that this type of frustration could be alleviated by dispensing more positive feedback in cases where a learner who has generally performed poorly takes the time to give a good response.

4.1.3 Segregating Participant Characteristics from Dialogue Features.

A critical hurdle that accompanies applications in the field of user modeling is whether observed patterns generalize across different participants (i.e., are they universal or person-specific?, Picard, 1997). In situations where relationships between patterns do not generalize, it is important to introduce some degree of normalization for each participant. The perceived scientific value of this research would increase if the interactions between the dialogue features and the learner's affective states would generalize above and beyond the participant's characteristics.

The multiple regression models described above were constructed in two steps, such that the dialogue features were excluded from the first step and were reintroduced in the second step. While this procedure quantitatively separates any explained variance into two groups, it fails to confirm whether any of the relationships observed in Table IV would reflect the commonality of variance between participant characteristics and dialogue features. That is, which of these two sources of variance should get credit when there is commonality of variance between the two?

One way to answer this question involved reversing the order in which the two sets of predictors are entered into the 2-step multiple regression analyses. If step-1 regression models are constructed on the basis of the 11 dialogue features only and the 29 participant characteristics variables are reserved for the step 2 model alone, one can determine whether differences among college students are significant above and beyond the dialogue features. Specifically, if a particular feature was a statistically significant predictor of an affective state in step 1 of the regression analyses (dialogue features only) as well as in step 2 (dialogue features plus participant characteristics) one could reliably conclude that any covariation observed between this feature and the affective state is valid and not caused by the participants' characteristics. On the other hand if a dialogue feature was a significant predictor in step 1 but not in step 2, then one would be forced to conclude that the relationship between the predictor and the affective states was a characteristic of individual learners and could not generalize beyond individual participants.

In accordance with this reverse regression procedure, the multiple regression analyses were repeated with the order of the dialogue features and participant characteristics reversed (dialogue features for model 1, dialogue features + participant characteristics for model 2). An examination of the significance and direction (+-) of the statistically significant predictors across both steps of the regression models did not reveal any serious contradictions to what was reported in Table IV. Therefore, we can conclude that our set of dialogue features does covary with the affective states above and beyond the participant characteristics.

4.2 Dimensionality Reduction

Dimensionality reduction is an important phase in machine learning experiments. In addition to potentially increasing classification accuracy by eliminating unrelated features, computational advantages are gained in terms of execution time. Therefore, the data were preprocessed before attempting to classify affect from dialogue. We pursued two methods of dimensionality reduction, one based on feature selection and the other based on feature extraction. The "feature selection first" method consisted of a supervised selection of features on the basis of the collinearity analyses and the multiple regression analyses described above. In particular the statistically significant standardized coefficients (after the elimination of 6 highly collinear features) listed in Table IV of the multiple regression analyses were used as the features for the classifiers.

The second dimensionality reduction technique involved extracting features by principal component analyses (PCA) and linear discriminant analysis (LDA). The PRAAT environment (Boersma & Weenink, 2006) was used to accomplish the requisite computation. The combination of these methods have been widely used and well validated as robust dimensionality reduction techniques, particularly in extracting features from speech.

Our analyses proceeded by first applying PCA to all 7 datasets, each containing the complete set of 17 features, and then dynamically reducing the dimensionality on the basis of the number of eigenvectors that accounted for 97% of the variance (typically 12-14). LDA was then applied to the decorrelated features to project them onto a lower dimensional space on the basis of the number of discriminant functions developed (number of classes – 1). Machine learning experiments were conducted, with the classifiers being trained on features extracted by the combined use of PCA and LDA as well as separately utilizing PCA and LDA. These strategies yielded classification accuracies that were slightly better than chance and are not discussed in the subsequent section. We suspect that this may be due to the fact that our set of predictors is quite small ($N = 17$) and hence extracting a subset may have reduced the predictive power.

4.3. Classifying Affective States from Conversation Features

In order to address the larger goal of extending AutoTutor into an affect-sensitive intelligent tutoring system, the need for real time automatic affect detection becomes paramount. Therefore, we applied 17 standard classification techniques in an attempt to detect the various affective states based on dialogue features. The motivation behind using a relatively larger set of classifiers was to determine which classifier yields the best performance. It also would be interesting to determine whether classifiers from any particular category (trees, rules, etc) outperforms the others.

The Waikato Environment for Knowledge Analysis (WEKA) (Witten & Frank, 2005) was used to comparatively evaluate the performance of various standard classification techniques in detecting affect from dialogue. The classification algorithms tested were selected from a list of categories including Bayesian classifiers (Naive Bayes and Naive Bayes Updatable), functions (Logistic Regression, Multilayer Perceptron, and Support Vector Machines), instance based techniques (Nearest Neighbor, K*, Locally Weighted Learning), meta classification schemes (AdaBoost, Bagging Predictors, Additive Logistic Regression), trees (C4.5 Decision Trees, Logistic Model Trees, REP Tree), and rules (Decision Tables, Nearest Neighbor Generalization, PART).

The classification process proceeded in three phases. In the first stage, we assessed the reliabilities of classifiers in discriminating between boredom, confusion, flow, frustration, and neutral. In the second phase, we only eliminated neutral and reduced the scope to boredom, confusion, flow, and frustration. This is a challenging task because the standardized coefficients of the regression models revealed that the diagnosticity of some of the dialogues features with respect to these four affective states was quite low. In the third phase of the classification analyses, we examined the accuracies of detecting each of the four affective states from the base state of neutral. Specifically, the classification algorithms were compared in their ability to differentiate boredom, confusion, flow, or frustration from neutral. There were challenges in this analysis as well. Although the multiple regression models' analyses provided a significant model for detecting the affective state of neutral, most of the variance associated with neutral was accounted for by the participants' pretest scores and overall emotion ratings; that is, the dialogue features were not very proficient predictors of neutral affect. The multiple regression analyses also failed to converge upon statistically significant models for delight and surprise, so these emotions were excluded from the classification analyses.

We established a uniform baseline for the different emotions by randomly sampling an equal number of observations from each affective state category. This sampling process was repeated for 10 iterations and all reported reliability statistics were averaged across these 10 iterations. For example, consider the task of detecting confusion from neutral with affect labels provided by the self. In this case we would randomly select an equal number of confusion and neutral samples, thus creating a data set with equal prior probabilities of both these emotions. Each randomly sampled data set was evaluated on the 17 classification algorithms and reliability statistics were obtained using k-fold cross-validation (k = 10).

A 3 factor repeated measures analysis of variance (ANOVA) was performed in order to comparatively evaluate the performance of the classifiers in detecting affect from the dialogue features. The first factor (*judge*) was the judge or combination of judges that provided the affect judgments. This factor had 7 levels: self, peer, trained judges 1, trained judge 2, trained judges agree, any 2 judges agree, and any 3 judges agree. The second factor involved the *emotions* classified and was composed of 6 levels: collectively discriminating between boredom, confusion, flow, frustration, and neutral (level 1, chance = 20%), discriminating between boredom, confusion, flow, and frustration (without neutral, level 2, chance = 25%), and individually detecting boredom, confusion, flow, and frustration from neutral (levels 3, 4, 5, and 6 respectively, chance = 50%). The third factor in the ANOVA was the classification scheme (called *classifier*) divided across 6 levels for Bayesian classifiers, functions, instance based learners, meta classifiers, rules, and trees. The unit of analysis for the 7x5x6 ANOVA was a single iteration of a single classifier. The kappa score was utilized as the metric to evaluate performance of each classifier because this metric partials out random guessing. The ANOVA indicated that there were significant differences in kappa scores across all three factors, as well as for various interactions between the factors. On the basis of the ANOVA we report comparisons between the various levels of our three factors (rater, emotion, and classifier). Figure 1 graphically depicts the mean kappa score obtained from the emotion classification for each level of each factor of the ANOVA.

4.3.1 Comparisons across Affect Judges. The results of the ANOVA indicated that a statistically significant effect was obtained for the judge $F(6,174) = 492.09$, $MSe = .009$, $p < .001$ (partial $\eta^2 = .944$). Bonferroni post hoc tests revealed that classifiers evaluated on data where at least three judges agreed ($M_{3A} = .295$, $p < .01$) yielded the best performance. However, this finding should be interpreted with caution since this data set probably consists of the most obvious cases, namely when three or more judges were able to agree on an affective state. It was also the smallest data set with only 410 records. The post hoc tests indicated that there were no significant differences in kappa scores for classifiers based on combined affect judgments where 2 or more judges agreed, the 2 trained judges agreed, and the judgments of trained judge 2 ($M_{2A} = .256$, $M_{JA} = .263$, and $M_{J2} = .258$). Classifiers trained on data with affect labels provided by trained judge 1 were lower ($M_{J1} = .245$) than these three scores. The lowest kappa scores were found in classifiers trained on affective judgments of the self ($M_{SF} = .111$, $p < .001$). Classifiers trained on affective judgments provided by the peer were significantly lower than all others with the exception of the self judgments ($M_{PR} = .162$, $p < .001$).

In general if one aggregates the 7 factors into 3 groups as novice judges (self and peer), trained judges (1 and 2), and combined models (trained judges agree, at least 2, and 3 judges agree) we obtain mean kappa scores of .137, .252, and .258. Therefore, one can conclude that reliability scores obtained by classifiers based on affect categories provided by trained judges and combined models were approximately the same and higher than those obtained by judgments provided by novice judges.

4.3.2 Comparisons across Emotions Classified. The ANOVA revealed statistically significant differences in kappa scores among the emotions classified $F(5,145) = 638.41$, $MSe = .013$, $p < .001$ (partial $\eta^2 = .957$). Bonferroni post hoc tests indicated that the classifiers were most successful in detecting frustration from neutral ($M_{FRNU} = .39$, $p < .001$). The classifiers had more success in collectively discriminating between boredom, confusion, flow, and frustration ($M_{WONU} = .229$) than individually detecting boredom, confusion, and flow from neutral ($M_{BONU} = .207$, $M_{CFNU} = .182$, $M_{FLNU} = .193$). Kappa scores obtained from efforts to detect boredom from neutral ($M_{BONU} = .207$) were significantly higher than similar efforts in detecting confusion from neutral ($M_{CFNU} = .182$).

The least robust results were obtained when we attempted to discriminate between the 5 affective states (boredom, confusion, flow, frustration, and neutral). In this case we obtained a mean kappa score of .163 (M_{ALL}), which was significantly lower ($p < .001$) than all other combinations of emotions. Discriminating a larger number of affective states is challenging, particularly when the states are collected in an ecologically valid setting (i.e. no actors were used to express emotions and no emotions were intentionally induced). Additionally, these results are on par with kappa scores associated with human judges (e.g. self-peer = .08, self-judge1 = .14, self-judge2 = .16).

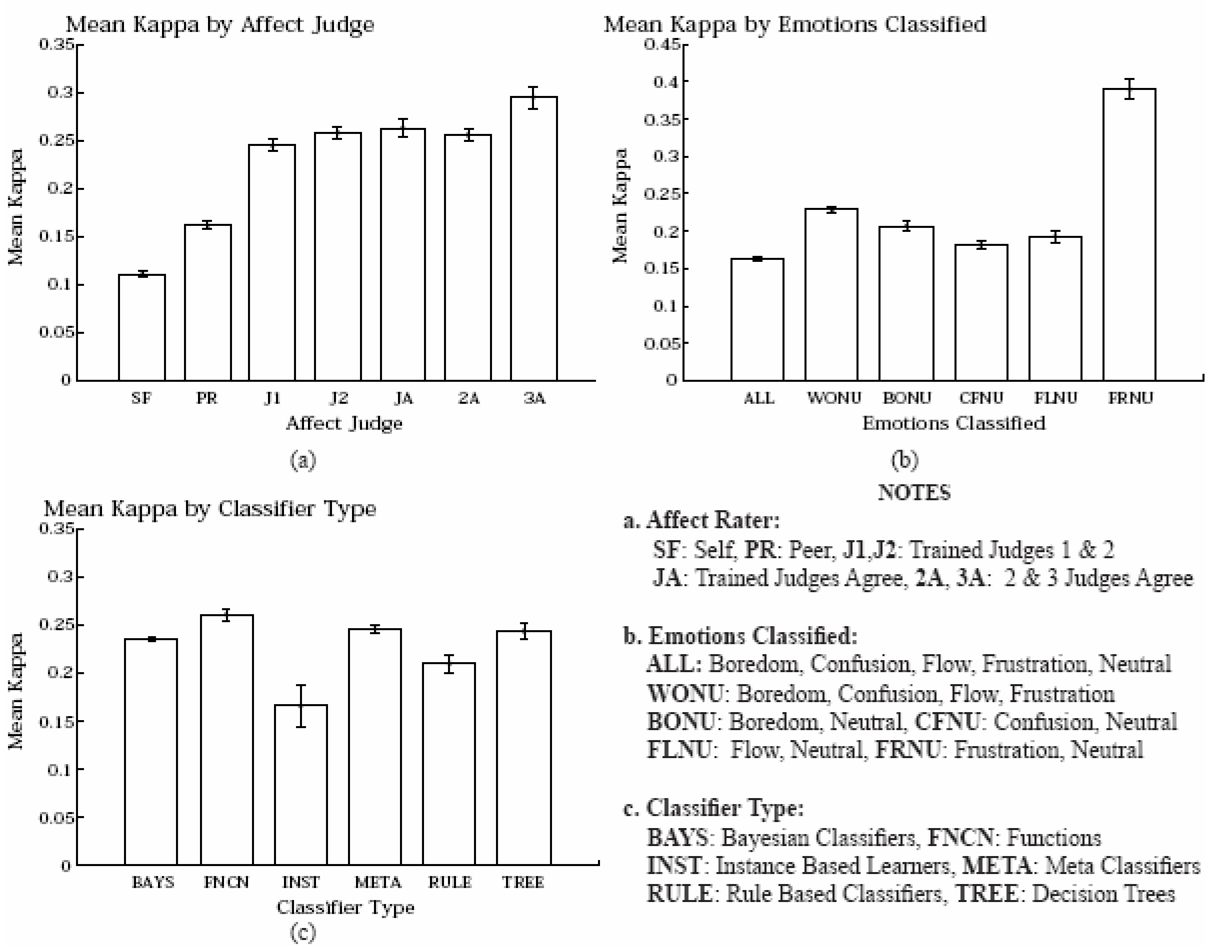

Figure 1. Mean kappa across: (a) Affect Judge; (b) Emotions Classified; (c) Classifier Type

4.3.3 Comparisons across Classifier Schemes. The results of the ANOVA indicated that there were statistically significant differences in the kappa scores across the various classifier schemes $F(5,145) = 44.83$, $MSe = .03$, $p < .001$ (partial $\eta^2 = .607$). Bonferroni post hoc tests revealed that the kappa scores of the functions, meta, and tree based classifiers ($M_{FNCN} = .261$, $M_{META} = .246$, $M_{TREE} = .244$) were similar quantitatively and significantly ($p < .01$) higher than the other categories. Kappa scores for the instance-based learning methods ($M_{INST} = .166$) were significantly lower than the other 5 classifier categories ($p < .001$ level). The post hoc tests also indicated that Bayesian classifiers ($M_{BAYS} = .236$) outperformed rule based classification schemes ($M_{RULE} = .21$, $p < .001$).

4.3.4 Optimal Classification Accuracy. The use of multiple assessments of the learner's affect (N=7) and a large number of classifiers (N=17) served an exploratory goal regarding the possibility of detecting affect from dialogue. However, in order to achieve our goal of developing a real time emotion classification system, we will shift our focus to the data set and classifier that yielded the best performance. Table V presents the maximum classification accuracies obtained across all 17 classifiers for each of the 7 data sets in collectively discriminating between the various affective states, as well as in individually detecting each state from neutral.

Table V. Comparison of various classification techniques to detect learner's affect

Affective States	Base Rate	**Max. Classification Accuracy (%)** Self	Peer	Judge 1	Judge 2	Judges Agree	2 Agree	3 Agree
Boredom,Confusion, Flow, Frustration, Neutral	20	29.5	31.4	38.3	*42.4*	*42.4*	41.4	42.2
Boredom,Confusion, Flow, Frustration	25	35.1	38.8	50.4	50.5	*54.0*	52.5	52.7
Boredom, Neutral	50	61.3	61.2	63.9	61.4	67.2	62.7	*69.0*
Confusion, Neutral		58.9	59.4	61.2	62.6	60.9	65.4	*67.8*
Flow, Neutral		52.9	56.0	66.8	70.0	*70.5*	63.9	67.3
Frustration, Neutral		64.1	69.2	73.5	76.7	76.6	76.0	*77.7*

Consider first the ability of classifiers to collectively discriminate boredom, confusion, flow and frustration, either with or without the neutral category being included. The best result, according to a simple logistic regression function, was provided by the trained judges and combined ratings. The accuracy was 42.4% and 54.0% with versus without neutral, respectively. The inclusion of neutral causes a reduction in accuracy, which is what could be expected since neutral is often confused with other emotions, particularly flow. This finding is consistent with the multiple regression models having difficulty in detecting neutral from these other emotions. The maximum accuracies in detecting boredom, confusion, and flow from neutral (69%, 68%, and 71%) were quantitatively similar and lower than accuracies involving the discrimination of frustration from neutral (78%). The AdaBoost classifier provided the best results in contrasting boredom and confusion from neutral. When discriminating neutral from flow and frustration, a simple logistic regression and C4.5 decision trees respectively yielded the best performance. With the exception of flow, the data sets that yielded the best performance in detecting boredom, confusion, and frustration from neutral were obtained from the data set in which 3 or more judges agreed on an affect rating. For flow and neutral, however, the best performance was obtained from the data set in which both trained judges agreed on the learner's affect. As with the multiple regression analyses, the classifiers operating on the judgments of individual judges (self and peer) were lower than those trained on the basis of the combined models, thus highlighting the merits of using composite affect judgments from multiple judges.

Table VI lists the F-measure scores for the affective states obtained from the most successful classifiers. The results indicate that when the classifiers attempted to collectively distinguish boredom, confusion, flow, and frustration, the reliabilities in detecting frustration and flow were similar and higher than those obtained for boredom and confusion. When one considers the analyses that distinguished each affective state from neutral, the affective state of frustration was most easily distinguished from neutral. The F-measure for the affective states of boredom, confusion, and flow were quantitatively similar.

Table VI. Accuracies for boredom, confusion, flow, and frustration, and neutral

Affective States	F-Measure						
	Self	Peer	Judge 1	Judge 2	Judges Agree	2 Agree	3 Agree
Boredom	.32	.38	.39	.29	.39	.39	.43
Confusion	.35	.21	.33	.43	.36	.42	.42
Flow	.13	.27	.47	.51	.52	.49	.51
Frustration	.35	.41	.48	.56	.56	.50	.47
Neutral	.26	.23	.16	.25	.23	.19	.23
Boredom	.35	.44	.45	.37	.47	.48	.52
Confusion	.43	.31	.44	.44	.44	.48	.49
Flow	.33	.41	.59	.59	.63	.60	.62
Frustration	.41	.49	.58	.59	.63	.60	.59
Boredom	.57	.68	.64	.63	.68	.64	.71
Neutral	.67	.61	.65	.62	.66	.67	.71
Confusion	.63	.64	.69	.66	.66	.67	.68
Neutral	.57	.59	.59	.63	.62	.66	.72
Flow	.59	.64	.68	.69	.70	.64	.72
Neutral	.54	.55	.66	.71	.71	.67	.70
Frustration	.66	.71	.75	.79	.79	.78	.80
Neutral	.63	.68	.72	.75	.75	.73	.75

4.3.5 Comparisons of Computer Generated Categories with Human Assessments. It is generally accepted that humans face some degree of difficulty in judging affect. The difficulty is apparent when we compare our assessments of emotions with other psychological entities. For example, methodologists in the social and behavioral sciences have sometimes claimed that kappa scores ranging from 0.4 – 0.6 are considered to be fair, 0.6 – 0.75 are good, and scores greater than 0.75 are excellent (Robson, 2003). However, emotions classification does not reach such a high bar of interrater agreement. Interrater reliability scores across a variety of research efforts involving emotion measurements by humans are in general quite low. For example, Litman and Forbes-Riley (2004) report kappa scores of around .4 in detecting positive, negative, and neutral affect. Ang et al. (2002) report a kappa score of .47 in human judgments of frustration and annoyance in human-computer dialogue. Shafran, Riley, and Mohri (2003) report kappa scores ranging from .32 to .42 in coding affect. Recently, Grimm et al., (2006) reported kappa scores of .48 for humans detecting acted emotions. We found that the highest kappa score obtained from our study was only .36. Therefore, an interesting research question is how well affect categories generated by the various automated classification algorithms compares with human classification of emotions, which has modest reliability at best.

In order to compare the reliability between computer generated emotion categories and human judgments, we conducted another set of classification analyses that focused on assessing reliabilities in collectively discriminating between boredom, confusion, flow and frustration. Since the previous analyses revealed that a simple logistic regression yielded the best performance in discriminating between these affective states, the subsequent analyses involve this classifier as representing the automated algorithms. In order to make a legitimate comparison with human judgments of affect, the combined data sets were not included in this analysis. Instead, classifiers were trained and evaluated on the four data sets that were constructed on the basis of individual judgments of the learner, a peer, and two trained judges.

In order to establish a uniform baseline, we sampled an equal number of observations for the affective states of boredom, confusion, flow, and frustration. This process was repeated for 10 iterations and accuracy results were averaged across these 10 iterations. After this sampling procedure, each of the four data sets (self, peer, 2 trained judges) were split into two parts: training data and testing data. The training data consisted of the dialogue features of 21 randomly sampled (without replacement) participants. The data for the remaining 7 participants served as the testing data. The logistic regression based classifier was then trained on the training data of a single judge and correspondingly evaluated on the testing data associated with the other three judges. For example, the classifier trained on the 21-participant subset of the self data was tested on the 7-participant subsets of data from the peer, and the 2 trained judges. In this manner four versions of the logistic regression classifier were constructed, each version being trained on the data of the self, the peer, and the two trained judges. Additionally, this entire process was repeated for 10 iterations, with each iteration involving the random sampling of a unique set of participants that constituted the training (75%) and testing (25%) data.

The results of the automated classifications are presented in Table VII, where they are contrasted with the judgments of the human coders. It should be noted that the reliability scores for the human coders differ from those reported earlier (i.e., self-peer = .08, self-judge1 = .14, self-judge2 = .16, peer-judge1 = .14, peer-judge2 = .18, and judge1-judge2 = .36). This is because the kappa scores for the human judges listed in Table VII were only performed on a sample of the affective states of boredom, confusion, flow, and frustration in order to make valid comparisons with the machine generated categories of learner emotions.

Table VII. Comparisons of computer generated affective states to human classification judgments

Human Measurements					**Computer Measurement**				
Rater 2	**Rater 1**				**Testing**				**Training**
	Self	*Peer*	*Judge 1*	*Judge 2*	*Self*	*Peer*	*Judge 1*	*Judge 2*	
Self	-	.131	.299	.288	-	.100	.257	.229	*Self*
Peer	-	-	.333	.343	.121	-	.278	.273	*Peer*
Judge 1	-	-	-	.583	.124	.169	-	.349	*Judge 1*
Judge 2	-	-	-	-	.102	.117	.266	-	*Judge 2*

A number of conclusions can be drawn from the results of the human-human (left of table) and computer-human interjudge reliability (right of table) scores presented in Table VII. Classifiers based on affect judgments of the self agreed with other human assessments of affect at a rate proportional to agreement between two humans (first row in Table VII). Human-computer agreement scores obtained when the classifiers were trained on the basis of the peer's affect judgments (second row) were slightly lower than human-human agreements for the two trained judges. Human-computer interjudge reliability scores obtained from classifiers trained on data provided by the two trained judges were significantly lower than the human-human interjudge reliability scores for the trained judges affect judgments (judge1-judge2$_{\text{human}}$ = .583 whereas judge1-judge2$_{\text{computer}}$ = .349, judge2-judge1$_{\text{computer}}$ = .266). On the basis of these observations, one can conclude that machine generated affect labels proportionally agree with novice judges (self and peer) but are inferior to trained judges.

5. Discussion

The problem of automating affect recognition is extremely challenging, on par with automating speech recognition. This project supports the conclusion that significant information can be obtained from AutoTutor's dialogue features, so dialogue can complement bodily measures for emotion detection. It appears that the classification accuracies obtained in this research on dialogue are not quite on par with the state-of-the-art algorithms that detect affect from facial features and speech contours. However, it should be noted that over a decade of sustained efforts have been directed towards affect detection from facial expressions and speech. This project is one of very few research investigations that classify affect from dialogue alone, whereas earlier efforts used a small number of dialogue features in conjunction with acoustic-prosodic and lexical features. Our results also constitute an improvement in classification accuracy compared to previous efforts that used dialogue features. We attribute this improvement primarily to the diversity and richness of our set of features. The features of dialogue in our analyses were specific to AutoTutor, but a similar set of features would presumably be relevant to any intelligent tutoring system, particularly in those that advocate deeper learning. In particular, the significant features that we extracted from AutoTutor's dialogue history logs (e.g., local good score, global good score, directness, etc.) would generalize to generic categories of dialogue features in all virtually all intelligent tutoring systems, such as content coverage, temporal parameters, response verbosity, student ability, tutor directness, and tutor feedback.

This section highlights contributions of this research towards the field of affect detection and human-computer interaction. We subsequently identify some of the limitations of this research and discuss improvements in affect classification that might mitigate these limitations and extend this line of research.

5.1 Research Overview

This paper has addressed three major research goals. These included (1) the collection of data on affect classification from multiple human judges and multiple channels, (2) statistical analyses that explore the relationship between a learner's affect and the various dialogue features, and (3) analyses of machine learning experiments that assess the accuracy of detecting affect from AutoTutor dialogue. This subsection summarizes and briefly speculates on the implications of our major findings.

5.1.1 Human Judgments of Learner Emotions. This multiple annotator study provided significant findings on human coding and online detection of affective states during learning with AutoTutor. The inter-rater reliability scores showed significant agreement on affective states among human raters, but there were informative differences among the raters. Judges trained in coding facial actions (Ekman & Friesen, 1978) showed comparatively high inter-judge agreement between themselves and matched the learner's self reports better than the untrained peers. This result suggests that untrained peers are not particularly adept at identifying the emotions of learners, whereas trained peers fare much better.

Models of affect that combine judgments made by the learner, a peer, and trained judges may represent an advance over traditional techniques that often rely on self reports of affect (De Vicente & Pain, 2002; Klein, Moon, & Picard, 2002; Matsubara & Nagamachi, 1996) or ratings by independent judges (Liscombe, Riccardi, & Hakkani-Tür, 2005; Litman & Forbes-Riley, 2004; Mota & Picard, 2003). The motivation behind these *composite* affect judgments of novice judges (self, peer) and trained judges resides in the indeterminacy of what exactly should be the gold standard for deciding what emotions a learner is truly having. In our study, would it be the learner or the expert on facial actions? A composite score that considers both viewpoints is arguably the most defensible position.

5.1.2 Dialogue-Emotion Links. The multiple regression analyses resulted in significant dialogue predictors for boredom, confusion, flow, frustration, and neutral but not for delight and surprise. The two-step multiple regressions allowed us to statistically partial out variance attributable to individual differences (which was a robust amount of variance) before assessing the unique impact of the conversation features on emotions. After partialling out individual differences, we found that the dialogue features were able to explain about 7% of the variance for boredom, confusion, flow, and frustration. We acknowledge that the explained variance is modest, but other researchers also report modest correlations and explained variance, as discussed throughout this article. Other researchers have not attempted to segregate the systematic variance that can be explained by individual differences per se versus intrinsic features of the dialogue. The generalizability of these results to other learners is supported by the significant relationships between the various dialogue features and the affective states that persisted after the removal of variables related to the individual learner characteristics. It is of course conceivable that there are hidden factors or interactions among the predictors that could explain additional variance between dialogue and affect, but tests of that possibility would require additional analyses.

It is also conceivable that the tutor-centered actions have a distinct influence on the affective states of the learner. Tutor-centered actions are moves and utterances of the tutor (feedback and directness) rather than the student (number of words in response, speech act, LSA measures, etc). We could explore this possibility by segregating the dialogue features listed in Table I into such categories as basic session information (subtopic and turn numbers), student-centered actions, and tutor-centered information. The amount of variance explained by each category of predictors could then be assessed by incrementally adding or removing each category in the multiple regressions analyses that predict affect categories. These analyses are planned in the future.

5.1.3 Automated Detection of Learner's Affect. The challenges of measuring emotions is beset with murky, noisy, and incomplete data, and is compounded with individual differences in experiencing and expressing emotions. Nevertheless we have found that the characteristics of the dialogue are quite diagnostic in predicting the affect states of learners. On the basis of the natural language dialogue features alone, our results showed that conventional classifiers are moderately successful in discriminating the affective states of boredom, confusion, flow, and frustration from each other, as well as from the baseline state of neutral.

The classification accuracy for collectively discriminating 5 affective states (including neutral) was significantly greater than the base rate. However, the reliability was lower compared to classifiers that individually detected each emotion from neutral or that collectively considered only the 4 emotions (excluding neutral). This motivates the use of a hierarchical classification scheme in order to improve accuracy. The hierarchical model would operate by first using a binary classifier to classify an incoming stimulus as a positive or a negative emotion, followed by an additional classification step that provides a finer discrimination as to what the individual positive or negative emotion may be (e.g., Hoque, Yeasin, & Louwerse, 2006). In a similar vein, we propose a hierarchical classifier motivated by a pandemonium model (Selfridge, 1959). A collection of affect-neutral classifiers 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 is declared as being experienced by the learner. If there is resonance with 2 or more emotions, then a second level of classification would be initiated in which classifiers collectively attempt to differentiate among boredom, confusion, flow, and frustration. Rigorous tests of these alternative hierarchical models will be pursued in future research.

5.2 Limitations

We acknowledge that our approach to data collection and analyses are not without limitations. This section reflects on a number of technical limitations and theoretical challenges.

5.2.1 Limited Contextual Information. One limitation of the data analyses presented in this paper is that each emotion judgment was analyzed in the context of only the immediately preceding turns of the student and tutor. The dialogue features that involved changed scores (delta local good, delta global good, delta local bad, and delta global bad) did encompass the context of one previous turn, but these features did not prove to be predictive of the affective states. Perhaps classification accuracies could be boosted by incorporating a broader scope of contextual information, including patterns of conversation that evolve over a series of turns leading up to an emotional experience. The exclusion of this larger snapshot of context preceding an emotion utterance could possibly account for some of the lower classifier accuracies. Future efforts will be directed towards the analysis of conversation features across a larger temporal span and number of turns.

5.2.2 Inability to Detect Delight and Surprise. The dialogue channels were unable to detect the affective states of delight and surprise. Perhaps these affective states are simply not manifested in AutoTutor's conversation features and their detection would require more sophisticated sensors. Delight and surprise are affective states that are generally expressed though animated facial features, so it may be possible to detect these states by means of the Facial Action Coding System (FACS, Ekman & Friesen, 1978). Ekman's research has associated action units 1 (inner brow raiser), 5 (raised upper eyelids), 26 (jaw drop), and 27 (mouth stretch) with surprise. While Ekman and Friesen (1978) did not investigate delight, they have associated action units with happiness, an emotion that is presumably similar to delight. In particular, action units 6 (raised lower eyelid), 7 (lid tightener), 12 (lip corner puller), 26 (jaw drop), and 27 (mouth stretch) have been affiliated with happiness. With the assistance of automated facial feature tracking software, we expect to be able to detect surprise and happiness (delight), thus compensating for the inability of detecting these affective states from AutoTutor dialogue.

5.2.3 Reliance on Shallow Assessments of Performance. A rather subtle limitation to the present results is that we relied exclusively on AutoTutor's assessments of the learner's contributions and its decisions regarding the type of feedback to give the student. Available research supports the claim that AutoTutor's assessments of the student's contributions highly correlate with human judgments (Graesser et al., 2007; Graesser et al., 2000; Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999) and that AutoTutor's conversational patterns have a close correspondence with human tutors (Person, Graesser, & TRG, 2002). However, AutoTutor's assessments are not error free or complete. There could be additions to the LSA scores in the computations of a learner's performance, such as measures of semantic entailment (Rus & Graesser, 2006) and the cohesion within each turn and across multiple turns. Cohesion refers to the linguistic properties of text that connect ideas conceptually. Perhaps student contributions with high cohesion may be indicative of a degree of understanding and would be diagnostic of the affective state of flow. On the other hand, a student contribution with low cohesion may be diagnostic of the affective state of confusion. A system called Coh-Metrix provides over 100 measures of various types of cohesion, including referential, spatial, temporal, causal, and structural cohesion (Graesser, McNamara, Louwerse, & Cai, 2004). Future efforts will be devoted to expanding the set of dialogue features to include measures of cohesion.

5.3 Future Directions

Considering the limitations mentioned above, we turn to a number of potential advances to this research. Some of these extensions directly address the larger project relating to AutoTutor's metamorphosis into an affect-sensitive intelligent tutoring system.

5.3.1 Combining Multiple Modalities. Affect detection in AutoTutor requires the development of appropriate classification systems. In addition to the conversation features and the learners' pretest knowledge, AutoTutor will be endowed with sensors that gauge facial expressions, speech intonation contours, and posture parameters. It will be interesting to determine whether classification performance from multiple channels will exhibit performance superior to an additive combination of the individual channels. An alternative possibility is that redundancy among channels may cause the multisensor classifier to yield negligible incremental gains.

We are in the initial phase of an investigation that attempts to isolate a subset of Facial Action Units (Ekman & Friesen, 1978) that are routinely observed in particular emotions during learning. In particular, Craig et al. (2004a) reported action units 1 (outer brow raise), 2 (inner brow raise), and 14 (dimpler) were primarily associated with frustration, with a strong link between action units 1 and 2 occurring together. Confusion displayed associations with action units 4 (brow lowerer), 7 (lid tightener), and 12 (lip corner puller); action unit 7 also triggered action unit 4. Boredom showed an association with 43 (eye closure). While boredom did not display any explicit links with the other action units, it did show several weaker trends between eye blinks and various mouth movements, such as mouth opening, mouth closing and jaw dropping, undoubtedly indicative of a yawn (Craig et al., 2004a).

Efforts towards affect detection from posture are motivated by the development of a Body Scoring System (Bull, 1987) and a study of interest detection in children (Mota & Picard, 2003). Detection of affective states from speech would involve the combination of acoustic and vocal prosodic features, as has been substantiated in previous emotion detection studies (Ang et al., 2002; Forbes-Riley & Litman, 2004; Litman & Forbes-Riley, 2004).

5.3.2 AutoTutor with Speech Recognition. Our multiple annotator study used a version of AutoTutor in which learner's typed in their contributions. Therefore, there were no data that related learner's affect with their speech patterns. Previous research has shown that acoustic-prosodic features of speech are excellent predictors of affective states, so we are in the process of systematically replicating the study with a newer version of AutoTutor in which the learner's verbally express their contributions. In addition to obtaining diagnostic data for affect detection from speech, we will attempt to replicate the various relationships between the dialogue patterns and the affective states discovered in this paper.

5.3.3 Use of Advanced Classifiers. The reliability of the standard classifiers in detecting affect from dialogue validates the notion of pursuing sophisticated classification techniques. Such techniques include biologically motivated classifiers as well as traditional classification methods. One biological approach, for example, is to develop classifiers that are based on the dynamic behaviors of neural populations (Kozma & Freeman, 2001). Classifiers based on dynamical systems with chaotic activity observed in brains have been experimentally validated as powerful pattern classifiers for difficult classification problems, particularly in situations in which the data set is not linearly separable (Kozma & Freeman, 2001).

6. Conclusion

The ultimate step in developing an affect sensitive AutoTutor lies in reengineering AutoTutor's dialogue planning module in a fashion that intelligently responds to the learner's affective states in addition to the cognitive states. This adaptation would increase the bandwidth of communication and allow AutoTutor to respond at a more sophisticated metacognitive level. There could be many possible responses to the different affective states of the learner and the context of the interaction. If the affective state of frustration is detected, then the ITS could respond by changing its dialogue strategies to include more direct feedback, assertions, and corrections of detected misconceptions. The tutor might also convey a degree of empathy to alleviate frustration. If the learner is bored, a state that has been negatively correlated with learning (Craig et al., 2004), then the ITS should engage the learner in a task that increases interest and cognitive arousal, such as a simulation, options of choice, a challenge, or a seductive embedded game. A change in dialogue strategies could be implemented to induce confusion by introducing related new topics or concepts, which will hopefully cause the student to reengage with the material at a deeper level. Confusion presents a key opportunity for the ITS to encourage deep learning. The AutoTutor system could manage confusion in at least two ways. Successful learners might be allowed to work out their own confusion in a discovery learning environment (Bruner, 1961; Vavik, 1993) that requires self-regulated cognitive activities (Azevedo & Cromley, 2004). A second method would systematically scaffold the student out of the confused state. This method might work better for learners with lower domain knowledge and lower ability to self-regulate their learning activities.

It is important to note that the success of the pedagogical strategies described above will ultimately depend upon the accuracy by which the learner's affect can be detected. The lower accuracies associated with detecting boredom and confusion are cause for some concern. We do expect that classification accuracy will increase when additional modalities are introduced. Nevertheless, it is unclear what a reasonable upper bound on emotion recognition accuracy would be. It may not be possible to achieve perfect accuracy, so a number of alternative methods may be recruited to handle emotion detection errors. AutoTutor could use probabilistic models, such as Dynamic Decision Networks, that can model the noisy data associated with recognizing emotions. AutoTutor could bias the confidence of the tutor's actions as a function of the confidence of the emotion estimate. For example if the ITS lacks confidence in its assessment of frustration, then an empathetic response may be preferred over AutoTutor's blatantly acknowledging the frustration and drastically altering its dialogue strategy.

Robust emotion detection is a significant challenge that must be solved to develop real-time, affect-sensitive tutoring systems that work. Only then will a learning environment that monitors learner emotions be more motivating and personally relevant to the learner. We hope that this research represents a small step towards fortifying future learners with ITS's capable of enhanced dynamic reasoning, automated cognitive assessment, and intelligent handling of emotions.

Acknowledgments

We thank our research colleagues in the Emotive Computing Group and the Tutoring Research Group (TRG) at the University of Memphis (http://emotion.autotutor.org). Special thanks to Barry Gholson, Jeremiah Sullins, Patrick Chipman, Max Louwerse, Kristy Tapp, Brandon King, and Stan Franklin for their valuable contributions to this study. We gratefully acknowledge our partners at the Affective Computing Research Group at MIT including Rosalind Picard, Ashish Kapoor, Barry Kort, and Robert Reilly.

This research was supported by the National Science Foundation (REC 0106965, ITR 0325428, and REC 0633918) and the DoD Multidisciplinary University Research Initiative administered by ONR under grant N00014-00-1-0600. 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, DoD, or ONR.

References

1. Aist, G., B. Kort, R. Reilly, J. Mostow and R. Picard: 2002, 'Adding human-provided emotional awareness to an automated reading tutor that listens'. *Intelligent Tutoring Systems 2002,* Berlin, Germany, pp. 992-993.
2. Aleven V. and K. R. Koedinger: 2002, 'An effective metacognitive strategy: Learning by doing and explaining with a computer based Cognitive Tutor'. *Cognitive Science* **26**, 147-179.
3. Alm, C.O. and R. Sproat: 2005, 'Perceptions of emotions in expressive storytelling'. *InterSpeech 2005*, Lisbon, Portugal, pp. 533-536.
4. Anderson, J. R., A. T. Corbett, K. R. Koedinger and R. Pelletier: 1995, 'Cognitive tutors: Lessons learned'. *The Journal of the Learning Sciences* **4**, 167-207.
5. Ang, J., R. Dhillon, A. Krupski, E. Shriberg and A. Stolcke: 2002, 'Prosody-based automatic detection of annoyance and frustration in human-computer dialog'. *Proceedings of the International Conference on Spoken Language Processing (ICSLP'02),* Denver, CO, pp. 2037-2039.
6. Azevedo, R. and J. G. Cromley: 2004, 'Does training on self-regulated learning facilitate students' learning with hypermedia'. *Journal of Educational Psychology* **96**, 523-535.
7. Batliner, A., K. Fischer, R. Huber, J. Spilker and E. Noth: 2003, 'How to find trouble in communication'. *Speech Communication* **40**, 117–143.
8. Bianchi-Berthouze, N. and C. L. Lisetti: 2002, 'Modeling Multimodal Expression of Users Affective Subjective Experience'. *User Modeling and User-Adapted Interaction* **12** (1), 49-84.
9. Boersma, P. and Weenink, D.: 2006, 'Praat: doing phonetics by computer (Version 4.3.14) [Computer program]'. Retrieved May 02, 2006, from http://www.praat.org/
10. Bosch, L. T.: 2003, 'Emotions, speech, and the ASR framework'. *Speech Communication* **40** (1-2), 213-215.
11. Bower, G. H.: 1981, 'Mood and Memory'. *American Psychologist* **36** (2), 129-148.
12. Bruner, J. S.: 1961, 'The act of discovery'. *Harvard Educational Review* **31** (1), 21-32.
13. Bull E. P.: 1987, 'Posture and Gesture'. Pergamon Press.
14. Campbell, D.T. and D. W. Fiske: 1959, 'Convergent and discriminant validation by the multitrait-multimethod matrix'. *Psychological Bulletin* 56, 81-105.
15. Carberry, S., L. Lambert and L. Schroeder: 2002, 'Toward Recognizing and Conveying an Attitude of Doubt Via Natural Language'. *Applied Artificial Intelligence* **16** (7), 495-517.
16. Cohn, J. F. and T. Kanade: In press, 'Use of automated facial image analysis for measurement of emotion expression'. In: J. A. Coan and J. B. Allen (eds.): *The handbook of emotion elicitation and assessment. Oxford University Press Series in Affective Science.* New York: Oxford.
17. Conati C.: 2002, 'Probabilistic assessment of user's emotions in educational games'. *Journal of Applied Artificial Intelligence* **16**, 555-575.

18. 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'. *Journal of Educational Media* **29**, 241-250.
19. Craig, S. D., S. D'Mello, A. Witherspoon, J. Sullins and A. C. Graesser: 2004a, 'Emotions during learning: The first step toward an affect sensitive intelligent tutoring system'. *Proceedings of the International Conference on eLearning,* Boston, MA: AACE, pp. 284-288.
20. Csikszentmihalyi, M.: 1990, 'Flow: The Psychology of Optimal Experience'. New York: Harper-Row.
21. De Vicente, A. and H. Pain: 2002, 'Informing the detection of students' motivational state: An empirical study'. *Intelligent tutoring systems 2002,* Berlin, Germany: Springer, pp. 933-943.
22. D'Mello, S. K., S. D. Craig, B. Gholson, S. Franklin, R. Picard and A. C. Graesser: 2005, 'Integrating affect sensors in an intelligent tutoring system'. *Affective Interactions: The Computer in the Affective Loop Workshop at 2005 International conference on Intelligent User Interfaces,* New York: AMC Press, pp. 7-13.
23. D'Mello, S. K., S. D. Craig, J. Sullins and A. C. Graesser: 2006, 'Predicting affective states through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue'. *International Journal of Artificial Intelligence in Education* **16**, 3-28.
24. Ekman, P and W. V. Friesen: 1978, 'The facial action coding system: A technique for the measurement of facial movement'. Palo Alto: Consulting Psychologists Press.
25. Ericsson, K. A. and H. A. Simon: 1993, 'Protocol analysis: Verbal reports as data'. Revised edition. Cambridge, MA: The MIT Press.
26. Forbes-Riley, K. and D. Litman: 2004, 'Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources'. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Boston, MA: Association for Computational Linguistics, pp. 201-208.
27. Fredrickson, B. L. and C. Branigan: 2005, 'Positive emotions broaden the scope of attention and thought-action repertoires'. *Cognition and Emotion* **19**, 313-332.
28. Gertner, A.S. and K. VanLehn: 2000, 'Andes: A coached problem solving environment for physics'. *Intelligent Tutoring Systems: 5th International Conference, ITS 2000,* New York: Springer, 133-142.
29. Goleman, D.: 1995, 'Emotional Intelligence'. Bantam Books: New York.
30. Gorin, A. L., G. Riccardi and J. H. Wright: 1997, 'How may I help you?'. *Speech Communication* **23**, 113–127.
31. Graesser, A. C., P. Wiemer-Hastings, K. Wiemer-Hastings, D. Harter, N. K. Person and Tutoring Research Group: 2000, 'Using latent semantic analysis to evaluate the contributions of students in AutoTutor'. *Interactive Learning Environments* **8**, 129-148.
32. Graesser, A., K. VanLehn, C. Rosé, P. Jordan, and D. Harter: 2001, 'Intelligent Tutoring Systems with Conversational Dialogue'. *AI Magazine* **22** (4), 39-51.
33. Graesser, A. C., N. Person, D. Harter and Tutoring Research Group: 2001, 'Teaching tactics and dialogue in AutoTutor'. *International Journal of Artificial Intelligence in Education* **12**, 257-279.
34. Graesser, A. C. and B. Olde: 2003, 'How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down'. *Journal of Educational Psychology* **95**, 524–536.
35. Graesser, A. C., D. S. McNamara, M. M. Louwerse and Z. Cai: 2004, 'Coh-Metrix: Analysis of text on cohesion and language'. *Behavioral Research Methods, Instruments, and Computers* **36**, 193-202
36. Graesser, A.C., P. Chipman, B. C. Haynes and A. Olney: 2005a. 'AutoTutor: An intelligent tutoring system with mixed-initiative dialogue'. *IEEE Transactions in Education* **48**, 612-618.
37. Graesser, A. C., N. Person, Z. Lu, M. G. Jeon and B. McDaniel: 2005b, 'Learning while holding a conversation with a computer'. In: L. PytlikZillig, M. Bodvarsson and R. Bruning (eds.): *Technology-based education: Bringing researchers and practitioners together,* Greenwich, CT: Information Age Publishing, pp. 143-167.
38. Graesser, A.C., B. McDaniel, P. Chipman, A. Witherspoon, S. D'Mello, and B. Gholson: 2006, 'Detection of emotions during learning with AutoTutor'. *Proceedings of the 28th Annual Conference of the Cognitive Science Society,* Mahwah, NJ: Erlbaum, pp. 285-290.
39. Graesser, A.C., P. Penumatsa, M. Ventura, Z. Cai, & X. Hu: 2007, 'Using LSA in AutoTutor: Learning through mixed initiative dialogue in natural language'. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*, Mahwah, NJ: Erlbaum., pp. 243-262.
40. Grimm, M., E. Mower, K. Kroschel and S. Narayan: 2006, 'Combining Categorical and Primitives-Based Emotion Recognition'. 14th European Signal Processing Conference (EUSIPCO), Florence, Italy.
41. Hoque, M. E., M. Yeasin and M. M. Louwerse: 2006, 'Robust Recognition of Emotion from Speech'. *IVA 2006, LNAI 4133*, Springer-Verlag Berlin Heidelberg 2006, pp. 42 – 53.

42. Hudlicka, E. and D. McNeese: 2002, 'Assessment of user affective and belief states for interface adaptation: Application to an Air Force pilot task'. *User Modeling and User-Adapted Interaction* **12** (1), 1-47.
43. Issroff, K. and T. del Soldato: 1996, 'Incorporating motivation into computer-supported collaborative learning'. *Proceedings of the European Conference on Artificial Intelligence in Education.*
44. Kim, Y.: 2005, 'Empathetic Virtual Peers Enhanced Learner Interest and Self-Efficacy'. *Workshop on Motivation and Affect in Educational Software at the 12th International Conference on Artificial Intelligence in Education.* Amsterdam, The Netherlands.
45. Klein, J., Y. Moon and R. Picard: 2002, 'This computer responds to user frustration – Theory, design, and results'. *Interacting with Computers* **14** (2), 119-140.
46. Koedinger, K. R., J. R. Anderson, W. H. Hadley and M. A. Mark: 1997, 'Intelligent tutoring goes to school in the big city'. *International Journal of Artificial Intelligence in Education* **8**, 30-43.
47. Kort, B., R. Reilly and R. Picard: 2001, 'An affective model of interplay between emotions and learning: Reengineering educational pedagogy—building a learning companion'. *Proceedings IEEE International Conference on Advanced Learning Technology: Issues, Achievements and Challenges,* Madison, Wisconsin: IEEE Computer Society, pp. 43-48.
48. Kozma, R. and W. J. Freeman: 2001, 'Chaotic Resonance: Methods and Applications for Robust Classification of Noisy and Variable Patterns'. *International Journal of Bifurcation and Chaos*, **11,** 1607-1629.
49. Landauer, T. K. and S. T. Dumais: 1997, 'A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge'. *Psychological Review* **104**, 211-240.
50. Landauer, T., D. McNamara, S. Dennis, and W. Kintsch (Eds.): In press, 'LSA: A road to meaning'. Mahwah, NJ: Erlbaum.
51. Lee, C. M. and S. Narayanan: 2004, 'Towards detecting emotions in spoken dialogs'. *IEEE Transactions on Speech and Audio Processing*.
52. Lepper. M.R. and R.W. Chabay: 1988, 'Socializing the intelligent tutor: Bringing empathy to computer tutors'. In: H. Mandl and A. Lesgold (eds.): *Learning Issues for Intelligent Tutoring Systems.* Hillsdale, NJ: Erlbaum, pp. 242-257.
53. Lepper, M. R. and M. Woolverton: 2002, 'The wisdom of practice: Lessons learned from the study of highly effective tutors'. In: J. Aronson (ed.): *Improving academic achievement: Impact of psychological factors on education.* Orlando, FL: Academic Press, pp. 135-158.
54. Lesgold, A., S. Lajoie, M. Bunzo and G. Eggan: 1992, 'SHERLOCK: A coached practice environment for an electronics troubleshooting job'. In: J. H. Larkin & R. W. Chabay (ed.): *Computer-assisted instruction and intelligent tutoring systems.* Hillsdale, NJ: Erlbaum, pp. 201-238.
55. Linnenbrink, E. and P. Pintrich: 2004, 'Role of Affect in Cognitive Processing in Academic Contexts'. In: D. Dai and R. Sternberg (ed.): *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development.* Mahwah, NJ: Lawrence Erlbaum.
56. Linnenbrink, E. A. and P. Pintrich: 2002, 'The role of motivational beliefs in conceptual change'. In: M. Limon and L. Mason (ed.): *Reconsidering conceptual change: Issues in theory and practice.* Dordretch, The Netherlands: Kluwer Academic Publishers, pp. 115-135.
57. Liscombe, J., G. Riccardi, D. and Hakkani-Tür: 2005, 'Using Context to Improve Emotion Detection in Spoken Dialog Systems'. *EUROSPEECH'05, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal.
58. Litman, D. J. and K. Forbes-Riley: 2004, 'Predicting student emotions in computer-human tutoring dialogues'. *Proceedings of the 42nd annual meeting of the association for computational linguistics*, East Stroudsburg, PA: Association for Computational Linguistics, pp. 352-359.
59. Litman, D. J., C. P. Rose, K. Forbes-Riley, K. VanLehn, D. Bhemhe, and S. Silliman: 2004, 'Spoken versus typed human and computer dialogue tutoring'. *Proceedings of the seventh international conference on intelligent tutoring systems*, Berlin: Springer Verlag, pp. 368-379.
60. Litman, D. J. and S. Silliman: 2004, 'ITSPOKE: An intelligent tutoring spoken dialogue system'. *Proceedings of the human language technology conference: 3rd meeting of the North American chapter of the association of computational linguistics*, Edmonton, Canada: ACL, pp. 52-54.
61. Mandler, G.: 1984, 'Mind and body: Psychology of emotion and stress'. New York: Norton.
62. Matsubara, Y. and M. Nagamachi: 1996, 'Motivation Systems and Motivation Models for Intelligent Tutoring'. *Proceedings of the Third International Conference in Intelligent Tutoring Systems.*
63. Miserandino, M.: 1996, 'Children who do well in school: Individual differences in perceived competence and autonomy in above-average children'. *Journal of Educational Psychology* **88**, 203-214.

64. Morimoto, C., D. Koons, A. Amir and M. Flickner: 1998, ‘Pupil Detection and Tracking using Multiple Light Sources’. Technical report, IBM Almaden Research Center.
65. Mota, S. and R. W. Picard: 2003, ‘Automated Posture Analysis for Detecting Learner’s Interest Level’. *Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction, CVPR HCI.*
66. Oliver, N., A. Pentland and F. Berand: 1997, ‘LAFTER: A Real-time Lips and Face Tracker with Facial Expression Recognition’. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* San Juan, Puerto Rico: IEEE, pp. 123-129.
67. Olney, A., M. Louwerse, E. Mathews, J. Marineau, H. Hite-Mitchell, A. Graesser: 2003, ‘Utterance Classification in AutoTutor’. *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications using Natural Language Processing*, pp. 1-8.
68. Ortony, A., G. L. Clore and A. Collins: 1988, ‘The Cognitive Structure of Emotions’. Cambridge, UK: Cambridge University Press.
69. Pantic, M. and L.J.M. Rothkrantz: 2003, ‘Towards an Affect-sensitive Multimodal Human-Computer Interaction‘. *Proceedings of the IEEE, Special Issue on Multimodal Human- Computer Interaction (HCI)*, vol. 91 (9), pp. 1370-1390.
70. Patrick B., E. Skinner and J. Connell: 1993, ‘What motivates children’s behavior and emotion? Joint effects of perceived control and autonomy in the academic domain’. *Journal of Personality and Social Psychology* **65**, 781-791.
71. Person, N.K., A. C. Graesser and Tutoring Research Group: 2002, ‘Human or computer?: AutoTutor in a bystander Turing test’. In: S. A. Cerri, G. Gouarderes, & F. Paraguacu (eds.): *Intelligent Tutoring Systems*. Berlin, Germany: Springer, pp. 821-830.
72. Picard, R.W.: 1997, ‘Affective Computing’. Boston, MA: MIT Press.
73. Picard, R.W., E. Vyzas and J. Healey: 2001, ‘Toward Machine Emotional Intelligence: Analysis of Affective Physiological State’. *IEEE Transactions Pattern Analysis and Machine Intelligence* **23** (10), 1175—1191.
74. Prendinger, H. and M. Ishizuka: 2005, ‘The Empathic Companion: A character-based interface that addresses users' affective states’. *International Journal of Applied Artificial Intelligence* **19** (3,4), 267-285.
75. Rani, P., N. Sarkar and C. A. Smith: 2003, ‘An Affect-Sensitive Human-Robot Cooperation – Theory and Experiments’. *Proceedings of the IEEE Conference on Robotics and Automation,* Taipei, Taiwan: IEEE, pp. 2382 – 2387.
76. Robson C.: 1993, ‘Real word research: A resource for social scientist and practitioner researchers’. Oxford: Blackwell.
77. Rus, V., & A. C. Graesser.: 2006, ‘Deeper natural language processing for evaluating student answers in intelligent tutoring systems’. *Proceedings of the American Association of Artificial Intelligence*. Menlo Park, CA: AAAI.
78. Russell, J. A.: 2003, ‘Core affect and the psychological construction of emotion’. *Psychological Review* **110**, 145-172.
79. Scheirer, J., R. Fernandez, J. Klein and R. Picard: 2002, ‘Frustrating the user on purpose: A step toward building an affective computer’. *Interacting with Computers* **14** (2), 93-118.
80. Schutzwohl A, and K. Borgstedt: 2005, ‘The processing of affectively valenced stimuli: The role of surprise’. *Cognition & Emotion* **19**, 583-600.
81. Selfridge, O. G.: 1959, ‘Pandemonium: A paradigm for learning’. *Symposium on the Mechanization of Thought Processes*, London: Her Majesty's Stationary Office, pp. 511-531.
82. Shafran, I. and M. Mohri: 2005, ‘A Comparison of Classifiers for Detecting Emotion from Speech’. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing,* Philadelphia, PA: IEEE, pp. 341-344.
83. Shafran, I., M. Riley and M. Mohri: 2003, ‘Voice signatures’. *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Piscataway, NJ: IEEE, pp. 31-36.
84. Silvia, P. and A. Abele: 2002, ‘Can positive affect induce self-focused attention? Methodological and measurement issues’. *Cognition and Emotion* **16**, 845-853.
85. Sleeman, D. and J. Brown: 1982, ‘Intelligent tutoring systems’. New York: Academic Press.
86. Stein, N. and L. Levine: 1991, ‘Making Sense Out of Emotion: The Representation and Use of Goal-structured Knowledge’. In: W. Kessen, A. Ortony, and F. Craik: *Memories, Thoughts and Emotions: Essays in Honor of George Mandler.* Hillsdale, NJ: Laurence Erlbaum Associates, pp. 295-322.
87. Tekscan: 1997, ‘Tekscan Body Pressure Measurement System User’s Manual’. South Boston, MA: Tekscan Inc.

88. VanLehn, K.: 1990, ‘Mind bugs: The origins of procedural misconceptions’. Cambridge, MA: MIT Press.
89. VanLehn, K., P. Jordan, C. P. Rosé, D. Bhembe, M. Bottner, A. Gaydos, et al.: 2002, ‘The architecture of Why2-Atlas: A coach for qualitative physics essay writing’. *Proceedings of the Sixth International Conference on Intelligent Tutoring,* Berlin: Springer – Verlag, 158-167.
90. VanLehn, K., A. C. Graesser, G. T. Jackson, P. Jordan, A. Olney and C. P. Rose: in press, ‘When are tutorial dialogues more effective than reading?’. *Cognitive Science.*
91. Vavik, L.: 1993, ‘Facilitating discovery learning in computer-based simulation learning environments’. In: R.D. Tennyson and A.E. Baron (eds.): *Automating instructional design: Computer-based development and delivery tools*. Berlin, Germany: Springer-Verlag, 403-449.
92. Vesanto, J., E. Alhoniemi, J. Himberg, K. Kiviluoto and J. Parviainen: 1999, ‘Self-Organizing Map for Data Mining in MATLAB: the SOM Toolbox’. *Simulation News Europe* **25**, 54, ARGE Simulation News, Vienna, Austria.
93. Walker, M.A., I. Langkilde-Geary, H. W. Hastie, J. Wright and A. Gorin: 2002, ‘Automatically training a problematic dialogue predictor for a spoken dialogue system’. *Journal of Artificial Intelligence Research* **16**, 293–319.
94. Whang, M. C., J. S. Lim and W. Boucsein: 2003, ‘Preparing Computers for Affective Communication: A Psychophysiological Concept and Preliminary Results’. *Human Factors* **45** (4), 623-634.
95. Wiemer-Hastings, P., K. Wiemer-Hastings and A. C. Graesser: 1999, ‘Improving an intelligent tutor's comprehension of students with latent semantic analysis’. In S.P. Lajoie and M. Vivet, Artificial Intelligence in Education, Amsterdam: IOS Press, pp. 535-542.
96. Witten, I. H. and E. Frank, 2005: ‘Data Mining: Practical machine learning tools and techniques (2nd ed.)’. San Francisco, CA: Morgan Kaufmann.

Table of Contents

List of Figures

List of Tables