Multimethod assessment of affective experience and expression during deep learning

Sidney K. D'Mello*

Department of Computer Science
University of Memphis
Memphis, TN 38152, USA
E-mail: sdmello@memphis.edu
*Corresponding author

Scotty D. Craig and Art C. Graesser

Department of Psychology
University of Memphis
Memphis, TN 38152, USA
E-mail: scraig@memphis.edu
E-mail: graesser@memphis.edu

Abstract: Inquiries into the link between affect and learning require robust methodologies to measure the learner's affective states. We describe two studies that utilised either an online or offline methodology to detect the affective states of a learner during a tutorial session with AutoTutor. The online study relied on self-reports for affect judgements, while the offline study considered the judgements by the learner, a peer and two trained judges. The studies also investigated the relationships between facial features, conversational cues and emotional expressions in an attempt to scaffold the development of computer algorithms to automatically detect learners' emotions. Both methodologies showed that boredom, confusion and frustration are the prominent affective states during learning with AutoTutor. For both methodologies, there were also some relationships involving patterns of facial activity and conversational cues that were diagnostic of emotional expressions.

Keywords: learning technology; affect; emotions; learning; dialogue; facial features; confusion; frustration; boredom; emote-aloud.

Reference to this paper should be made as follows: D'Mello, S.K., Craig, S.D. and Graesser, A.C. (2009) 'Multimethod assessment of affective experience and expression during deep learning', *Int. J. Learning Technology*, Vol. 4, Nos. 3/4, pp.165–187.

Biographical notes: Sidney K. D'Mello is a Researcher at the Institute for Intelligent Systems at the University of Memphis, USA. He is active in a number of research areas including affective computing, intelligent tutoring systems, spoken dialogue systems, computational models of human cognition and radio frequency identification. He has authored 10 journal articles, 6 book chapters, 41 conference and workshop proceedings, and 16 non-published conference presentations in these and related research areas. D'Mello received a BS in Electrical Engineering from Christian Brothers University and an MS in Mathematical Sciences from the University of Memphis. He is currently pursuing a PhD in Computer Science at the same university.

Dr. Scotty D. Craig is a Research Scientist at the University of Memphis with the Institute for Intelligent Systems/Department of Psychology. To date, he has worked on projects in such areas as affect and learning, discourse processing, multimedia learning, vicarious learning environments and intelligent tutoring systems in both laboratory and applied classroom settings. His contributions to the field to date consist of 13 journal articles, 1 book chapter, 3 manuscripts in press, 30 published conference proceedings, 32 non-published conference presentations and 2 manuscripts submitted/in progress. Dr. Craig has served as an ad hoc reviewer for ten academic journals and conferences. He has organised two individual workshops and served as workshop chair for AIED 2009. He is currently on the editorial board of the *International Journal of Learning Technology*.

Art C. Graesser is a Professor in the Department of Psychology and Computer Science and the Co-director of the Institute for Intelligent Systems at The University of Memphis, USA. Dr. Graesser received his BA in Psychology from Florida State University and his PhD in Psychology from the University of California at San Diego. Dr. Graesser has worked in several areas of cognitive science, artificial intelligence and discourse processing, including text comprehension, inference generation, conversation, question asking and answering, tutoring and advanced learning environments (including AutoTutor). He has authored over 400 technical articles, authored 2 books and edited 8 books.

1 Introduction

In recent years, monitoring of human emotions or affective states during deep learning of complex topics has moved out of its infancy into a more robust area of research. Deep learning involves the learner generating explanations, justifications, and functional procedures instead of reading information and memorising definitions, facts, and properties (*i.e.*, shallow learning) (Graesser *et al.*, 2007b). The recent interest in the link between affect and learning is based on the assumption that deep learning is not entirely limited to cognition, discourse, action, and the environment because emotions (affective states) are inextricably bound to the learning process (Lepper and Henderlong, 2000; Linnenbrink and Pintrich, 2002; Meyer and Turner, 2006; Stein and Hernandez, in press). A prediction stemming from this assumption is that an agile learning environment that is sensitive to a learner's affective states presumably enriches learning, particularly when deep learning is accompanied by confusion, frustration, boredom, interest, excitement, and insight (D'Mello *et al.*, 2007; Graesser *et al.*, 2007a; Kort *et al.*, 2001; Lepper and Chabay, 1988; Lepper and Woolverton, 2002; Picard, 1997).

Notable among the recent research activities investigating the link between emotions and learning are efforts to incorporate assessments of learners' affect into the pedagogical strategies of Intelligent Tutoring Systems (ITSs) and peer learning companions (*e.g.*, Conati, 2002; Kort *et al.*, 2001; Litman and Forbes-Riley, 2004; McQuiggan and Lester, 2007; Woolf *et al.*, 2007). For example, Kort *et al.* (2001) proposed a comprehensive four-quadrant model that explicitly links learning and affective states. This model was used in the MIT group's work on their *affective learning companion*, a fully automated computer program that recognises a learner's affect by monitoring facial features, posture

patterns, and onscreen keyboard/mouse behaviours. Taking a slightly different approach, Conati (2002) developed a probabilistic system that can track multiple emotions of the learner during interactions with an educational game. Her system relies on dynamic decision networks to assess the affective states of joy, distress, admiration, and reproach. Litman and Forbes-Riley (2004) work with their ITSPOKE conceptual physics ITS uses a combination of discourse markers and acoustic-prosodic cues to detect and respond to a learner's affective states. Recently, Graesser and colleagues have been integrating affect sensing devices that monitor facial features, body position and movement, speech contours, and discourse features into AutoTutor, a natural language ITS (see *A Brief Overview of AutoTutor* section). The goal of the project is to endow AutoTutor with the ability to be responsive to the affective and cognitive states of a learner (D'Mello *et al.*, 2005; 2007).

The systems above are motivated by the assumption that there is an inextricable link between emotion and cognition. A better understanding of affect-learning connections is needed to design engaging educational artefacts that range from affect-sensitive ITSs on technical material (D'Mello *et al.*, 2005; 2007; Kort *et al.*, 2001; Litman and Forbes-Riley, 2004; Woolf *et al.*, 2007) to entertaining media and serious games (Conati, 2002; Gee, 2003; McQuiggan and Lester, 2007; Vorderer, 2003). However, emotions are notoriously difficult to study due to inherent variations across personalities, experience, age, gender, culture, and time. The field therefore needs a systematic multifaceted way to explore the connections between affective states and complex learning. This paper addresses this goal by identifying some of the important states that occur during learning, comparing methodologies for monitoring these affective states and developing systems to automatically detect the affective states in real time.

1.1 Goal 1. Identifying the affective states that accompany complex learning

There have been theories that link cognition and affect very generally, such as those of Bower (1981), Mandler (1999), Ortony *et al.* (1988), Russell (2003), and the more recent one by Stein and Hernandez (in press). While these theories convey general links between cognition and emotions, they do not directly explain and predict the emotions that occur during complex learning, such as attempts to master physics, biology, or computer literacy. Some emotions undoubtedly have a more salient role in learning than others (Linnenbrink and Pintrich, 2002). Researchers in different fields are familiar with Ekman's pioneering work on the detection of emotions from facial expressions (Ekman and Friesen, 1978). However, the emotions that Ekman intensely investigated (sadness, happiness, anger, fear, disgust, surprise), though ubiquitous to everyday experience, have minimal theoretical relevance to learning (Kort *et al.*, 2001) and do not tend to be found in studies that have investigated affect during 1–2 h learning sessions (D'Mello *et al.*, 2006; Graesser *et al.*, 2007a; Lehman *et al.*, 2008; Pekrun *et al.*, 2002). Researchers have proposed a different set of emotions that prevail during complex learning, namely boredom (Miserandino, 1996), confusion (Craig *et al.*, 2004; Graesser *et al.*, 2005; Kort *et al.*, 2001), frustration (Kort *et al.*, 2001; Patrick *et al.*, 1993), delight (Fredrickson and Branigan, 2005; Silvia and Abele, 2002), and flow (Csikszentmihalyi, 1990). Empirical tests of this list of learning-centred affective states is the first goal of our research.

1.2 Goal 2. Comparing methodologies to monitor affective states

Although automated affect detection systems are on the horizon, the majority of affect research still relies on humans to measure affect. This raises some interesting methodological questions about the identity of the person making the judgement (self reports or external observers) and the time (online or offline) of the judgement. As with every design decision, these alternatives have tradeoffs that need to be carefully evaluated. For example, one fairly common method for determining emotions from humans is to use a *self report questionnaire* administered after the experimental stimuli is presented (Larsen *et al.*, 2001). These measures are limited by the reporter's ability and sensitivity to one's emotions, as well as the reporter's ability to be honest. The judgements may also be influenced by other off-line measures at the time of testing. Thayer (1989) reported that this type of self report method for arousal showed a weak relationship to other indexes of autonomic arousal, so it is questionable whether self report questionnaires provide valid results in the context of learning.

Another measure involves human judges providing affect judgements while observing participants in a learning session. For example, Craig *et al.* (2004) conducted an online observational study in which participants' affective states were coded by observers during interactions with AutoTutor. The results revealed that learning gains were positively correlated with flow/interest and confusion, negatively correlated with boredom, and uncorrelated with frustration, eureka, and neutral states. The correlation between confusion and learning is perhaps counterintuitive, but confusion is undoubtedly affiliated with experiencing impasses, breakdown scenarios, and the resulting deep thinking (Graesser *et al.*, 2005; VanLehn *et al.*, 2003). Although the Craig *et al.* (2004) study provided interesting results, the observational coding method suffers from three potential problems. The first is similar to problems found with offline self reports. The affect measurements are based upon an observer of the learning session, so they are reliant upon the attentiveness, sensitivity, and expertise of the observer. Second, the observations were not recorded so there was no option of reinspecting the behaviours at a later time. Additionally, there were no self reports or additional measures that the observed judgements of affect could be correlated with. The third potential problem with this method is that it is highly reactive if learners know that a person is actively observing their affective states. This could very well cause them to exaggerate positive affect states or suppress the display of negative states. Therefore, assessing the impact of different methodologies on affect measurement is an important, second goal of this paper.

1.3 Goal 3. Automated detection of learning-centred emotions

Affect sensitive interfaces are guided by the design goal of narrowing the gap between the emotionally challenged computer and the emotionally rich human. Robust recognition of emotions is a critical requirement for such systems because expectations are raised when humans recognise that a computer system is attempting to communicate at their level (*i.e.*, with enhanced cognitive and emotional intelligence). When these expectations are not met, users often get discouraged, disappointed, or even frustrated (Norman, 1994; Shneiderman and Plaisant, 2005). Although, affect recognition accuracy need not be perfect, but it should be approximately and sufficiently on target.

The development of an automated affect-recognition system is a challenging endeavour because emotional expressions are sometimes murky, subtle, and compounded with individual differences in experience and expression. Therefore, robust recognition of users' emotions is a crucial challenge that is hindering major progress towards the larger goal of developing affect-sensitive interfaces that work.

Recently there have been ground-breaking advances in computational systems that classify human emotions (see Pantic and Rothkrantz, 2003 for a comprehensive review and Paiva *et al.*, 2007 for recent updates). Most of these affect-detection systems attempt to recognise Ekman and Friesen's (1978) basic emotions (anger, fear, sadness, happiness, disgust, and surprise). As discussed earlier, however, these basic emotions are not particularly relevant to learning. Therefore, there is a need for computational systems to detect the presence of some of the learning-centred emotions (confusion, frustration, boredom, flow/engagement, delight, and surprise). Progress in achieving the primary goal requires an interdisciplinary integration of computer science, psychology, artificial intelligence, and artefact design. This paper addresses this goal by identifying some of the cognitive and bodily correlates of affective experience and discussing how this information can be used to develop automated affect detection systems.

1.4 Overview of paper

In this paper, we describe two studies that attempted to address the three goals discussed above. Both studies involved humans monitoring the affective states that learners experienced during interactions with AutoTutor. However, the studies used different methodologies on the same population (*i.e.*, college students) of learners. Study 1 had an emote-aloud protocol in which *learners* verbalised their emotions *while* interacting with the tutor (D'Mello *et al.*, 2006). In Study 2, learners' affect was measured by *multiple judges* (*i.e.*, the learner, an untrained peer, and two trained judges) via a retrospective affect judgement procedure that occurred *after* the learner interacted with the tutor (Graesser *et al.*, 2006). The differences between the *affect judges* (learners themselves versus learners + other judges) and the *time* of the measurement (online vs. offline) constitute the major differences between the two methodologies.

We addressed the goal of obtaining a set of representative emotions that accompany learning (Goal 1) by monitoring both basic emotions and learning-centred affective states. Progress towards the goal of developing automated affect detection systems (Goal 3) was investigated by monitoring the cognitive and bodily correlates of emotional expression. The cognitive correlates included conversational cues and dialogue features obtained from AutoTutor's natural language mixed-initiative dialogue. The bodily correlates of emotional expression were monitored by facial feature tracking.

Goal 2 was implicitly investigated by virtue of the fact that two different methodologies were used to identify the affective states that accompany learning, along with their cognitive and bodily correlates. The use of different methodologies on the same population to investigate Goals 1 and 2 has two unique advantages. First, it allows us to explore whether any of the findings in Study 1 generalise to Study 2. This can be accomplished by analysing each study independently, identifying reliable patterns, and assessing whether these patterns are observed across studies. Any patterns that do generalise can be attributed to an inherent characteristic of affect-learning interactions and not an artefact of the methodology. The second advantage is that some of the

differences observed across studies can be attributed to methodological factors. However, as opposed to systematic replications, where a single factor is varied across studies, the two studies varied along a number of factors. Although, this reduces our ability to make causal inferences on the impact of any given factor that varied across studies, any differences observed can become testable hypotheses for further research.

The paper is organised in five sections. First, we provide a brief overview of AutoTutor, which was the learning environment in both studies. The next two sections describe the emote-aloud and multiple judge studies. The results section identifies a set of affective states that were observed in both studies, along with a generalisable set of their cognitive and bodily correlates. Finally, the general discussion section provides some of the advantages and disadvantages of each methodology. We also discuss the prospects of developing automated systems to identify some of the more prominent learning-centred affective states.

2 A brief overview of AutoTutor

AutoTutor is an Intelligent Tutoring System that helps learners construct explanations by (a) interacting with them in natural language and/or (b) helping them use simulation environments (Graesser *et al.*, 2005; 2001). AutoTutor attempts to comprehend students' natural language contributions and then responds to students' typed or spoken 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 *summarises* topics. During the tutorial dialogue, AutoTutor attempts to elicit information from the learner by first providing hints, then prompts and finally stating the missing information to the learning via assertions. 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.

The impact of AutoTutor in facilitating the learning of deep conceptual knowledge has been validated in over a dozen experiments on college students as learners for topics in introductory computer literacy (Graesser *et al.*, 2004), conceptual physics (VanLehn *et al.*, 2007), and critical reasoning on scientific methods (Storey *et al.*, in press). Tests of AutoTutor have produced gains of .4 to 1.5 sigma (a mean of .8), depending on the learning measure, the comparison condition, the subject matter, and version of AutoTutor. It should be pointed out that the amount of training time and the number of AutoTutor questions covered in the studies reported here was much less than previous tutoring sessions with AutoTutor that systematically assessed learning gains (Graesser *et al.*, 2004; VanLehn *et al.*, 2007). Moreover, the goals of the present study were to analyse emotions during learning from AutoTutor rather than assessing learning gains. Given the short amount of training time and the small number of questions covered, we did not expect impressive learning gains and therefore did not perform

systematic analyses on relationships between learning and emotions. From the standpoint of this paper, we will take it as given that AutoTutor helps deep learning whereas our direct focus is on the emotions that accompany the learning process.

3 Emote-aloud during learning with AutoTutor (Study 1)

We implemented an emote-aloud protocol that allowed for the implementation of online self reports in a real time setting (Craig *et al.*, 2008; D'Mello *et al.*, 2006). Although, this method is based on the subjective report of the learner, we hypothesise that the real-time reporting of the affective states will give a more reliable measure of what states occurred and when they occurred. This will reduce the impact that later events at subsequent testing could have on perceived affective experience. Of course, as with all methodologies involving human measurement of emotions, participants should possess the requisite degree of emotional intelligence (Goleman, 1997) to be able to accurately report their emotions.

The emote-aloud procedure is a modification of the think-aloud procedure (Ericsson and Simon, 1993). When think-aloud protocols are collected, participants talk about their thought process while working on tasks that require deeper levels of thought, such as solving problems (Ericsson and Simon, 1993), comprehending text (Trabasso and Magliano, 1996), or reading poetry (Eva-Wood, 2004). Our emote-aloud procedure works in a similar way. Participants were asked to simply state the affective states they were feeling while learning about computer literacy with AutoTutor. This method allows for online identification of emotions while working on a task with minimal task interference.

Think-aloud studies and this emote-aloud study collect data from a small number of participants because of the labour-intensive nature of the data collection and analysis (*e.g.*, transcription of protocols, segmenting and identifying meaningful units, scoring interjudge reliability). For example, Newell and Simon's (1972) pioneering work on problem solving had less than a handful of participants contributing think aloud data. Chi *et al.*'s (1989) classical work on self-explanation similarly had a small sample of participants. The number of participants can be small, yet still yield rich and reliable data (Ekman, 2003). Furthermore, concerns stemming from our ability to generalise from a small number of participants can be alleviated by assessing whether any patterns discovered are replicated in the study with multiple judges and a larger sample of learners (Study 2).

3.1 Brief sketch of methodology

3.1.1 Measuring affective states

The emote-aloud study had seven learners who were videotaped while interacting with the AutoTutor system for approximately 1.5 h. The learners were given a list of affective states and definitions (see Appendix). They were instructed to speak aloud particular emotions whenever they experienced one of eight affective states: anger, boredom, confusion, contempt, curious, disgust, eureka, and frustration (see D'Mello *et al.*, 2006 for a detailed description of the methodology). However, the emote-aloud procedure

produced a sufficient number of observations only for boredom, confusion, eureka, and frustration. The other affective states were reported infrequently, if at all, as will be discussed later.

3.1.2 Coding facial expressions

Ekman and Friesen's (1978) Facial Action Coding System (FACS) was adopted when we analysed the facial features that accompanied the various emotions. FACS specifies how judges are to code specific facial behaviours (*i.e.*, *action unit* or AU), based on the muscles that produce them. Facial expressions of affective states tend to be quite short, rarely lasting for more than three seconds (Ekman and Friesen, 1978). Therefore, two coders independently scored the three seconds before an emote-aloud report was made using FACS (see Craig *et al.*, 2008 for more details).

3.1.3 Coding dialogue features

A session with AutoTutor involves the student and the tutor collaboratively working on a solution to a problem. A large proportion of this collaborative problem solving process is realised by three-step dialogue cycles:

Step 1 AutoTutor asks a question (*i.e.*, hints or prompts) or asserts some information.

Step 2 The student provides a response.

Step 3 AutoTutor evaluates the response and provides feedback.

We mined three features from AutoTutor's log files in order to explore the links between the different phases in this cycle and the affective states of the learners. The first feature was the type of dialogue move that AutoTutor used to implement Step 1 of the cycle. This move was ordered on a scale on the basis of the amount of information AutoTutor supplies to the learner. The ordering of this *tutor directness* scale is pump (–1.0) < hint (–0.5) < prompt (0.0) < assertion (0.5) < summary (1.0). AutoTutor's pump (*e.g.*, 'what else?', 'tell me more') conveys the minimum amount of information (on the part of AutoTutor) whereas a summary conveys the most amount of explicit information. Tutor directness was expected to have an impact on emotions to the extent that pumps, hints (*e.g.*, 'what about X'), and prompts (*e.g.*, 'X is a type of __') create uncertainty in the mind of the student.

The second feature was the conceptual quality (*answer quality*) of students' responses, as measured by Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997) and other semantic components. The answer quality measure was the semantic match between the learner's response and the expected answer. Students with higher quality answers are expected to be performing better and thereby experiencing more positive emotions.

The third feature was the type of feedback AutoTutor provided to the learner. The levels of the *tutor feedback* scale were negative feedback (–1.0, *e.g.*, 'wrong', 'no'), neutral-negative feedback (–0.5, *e.g.*, 'possibly', 'kind of'), neutral feedback (0, *e.g.*, 'uh huh', 'alright'), neutral-positive feedback (0.5, *e.g.*, 'yeah', 'hmm right'), and positive feedback (1.0, *e.g.*, 'good job', 'correct'). The student's emotions are expected to be systematically influenced by the feedback in the obvious direction: positive emotions after positive feedback and negative emotions after negative feedback.

4 Offline emotion judgements by multiple judges (Study 2)

As in the emote-aloud study, researchers sometimes have relied on a single operational measure in inferring a learner's emotion, such as self reports (De Vicente and Pain, 2002; Klein *et al.*, 2002) or ratings by independent judges (Litman and Forbes-Riley, 2004; Mota and Picard, 2003). However, as mentioned above, the accuracy of self reports in measuring affect is not clearly understood (Russell, 2003; Thayer, 1989). Furthermore, there are no conclusive reasons to expect it to be extensively high (Graesser *et al.*, 2006).

We employed an offline methodology with multiple raters, the learner, a peer, and two trained judges. Employing multiple measures of affect is compatible with the standard criterion for establishing convergent validity (Campbell and Fiske, 1959). This methodology overcomes some measurement problems by including multiple raters to lessen the impact of observer bias. The session was also recorded so that raters could make more thoughtful reviews after the learner had completed their tutorial session with AutoTutor. The affective states included in this study were frustration, confusion, flow, delight, surprise, boredom, and neutral. Contempt, curious, disgust, and anger were excluded from this study because they rarely occurred in the emote-aloud study. Delight and surprise were added as functional replacements for eureka because of validity concerns associated with eureka (described below). Flow was not included in the emote-aloud study due to a concern that requiring participants to emote-aloud on their flow experiences would disrupt the flow experience. However, flow was included in the list of emotions in the multiple-judge study because affect judgements occurred after the interaction session. Finally, neutral was added because participants were required to make forced-choice affect judgements.

4.1 Brief sketch of methodology

Study 2 had 28 college students who interacted with the AutoTutor system for 35 min. A video of the learner's face and a video of the computer screen was recorded. The judging process was initiated by synchronising the video streams from the screen and the face, and displaying them to the judges. Judges were instructed to make judgements on what affective states were present in 20-sec intervals (*mandatory* judgements), at which time the video automatically paused. They were also instructed to indicate any affective states that were present in between the 20-sec stops (*voluntary* judgements). Mandatory and voluntary judgements occurred within the same judging session and judges were able to replay the 20 sec segment prior to making an emotion judgement. The sampling rate for this methodology is consistent with the 'thin slices' idea of Ambady and Rosenthal (1992), where it has been shown that raters can determine affect by observing brief clips of video sometimes as briefly as 2.5 sec long (Strahan and Zytowski, 1976). Our consistent sampling rate of 20 sec was sufficiently long to identify the learner's affect and also allowed ample time for valid voluntary judgements to occur.

Four sets of emotion judgements were made for each learner's AutoTutor session. For the self judgements, learners watched their own sessions with AutoTutor immediately after having interacted with the tutor. For the peer judgements, the participants returned a week later to watch and judge another learner's session on the same topic in computer literacy (*i.e.*, operating systems, hardware, or the internet). Finally, two additional trained judges analysed all of the sessions separately. These trained judges

had been trained on how to detect facial action units according to Ekman's Facial Action Coding System (FACS) (Ekman and Friesen, 1978). The trained judges also had considerable experience with AutoTutor and used their knowledge of AutoTutor's dialogue characteristics (*i.e.*, context) along with their expertise in detecting facial features in making their judgements.

4.1.1 Coding facial expressions and dialogue features

The process of coding of facial and dialogue features was the same as the emote-aloud study.

5 Results and discussion

5.1 The incidence of affective states (Goal 1)

5.1.1 Emote-aloud study

In the emote-aloud study, there was a significant difference in the proportions of the affective states produced by learners in the emote-aloud task, $F(7, 42) = 7.90$, $Mse = .011$, $p < .001$. Table 1 shows frequencies, mean proportions, and standard deviations for the various emotions in both Study 1 and Study 2. It appears that frustration, boredom, confusion, and eureka represent the majority of the emotions experienced by learners, accounting for 85.3% of the self reports. Occurrences of contempt, anger, disgust, and curious were much rarer, collectively comprising a meager 14.7 of the emotional experiences.

Table 1 Proportion of affective states observed

	Emote-aloud study (online)			*Multiple judge study (offline)*		
		Proportions			*Proportions*	
Affective states	*Frequency*	*Mean*	*Stdev*	*Frequency*	*Mean*	*Stdev*
Anger	17	.054	.024	–	–	–
Boredom	43	.260	.052	553	.249	.126
Confusion	54	.171	.044	835	.354	.115
Contempt	8	.061	.031	–	–	–
Curious	1	.003	.003	–	–	–
Delight	–	–	–	112	.052	.058
Disgust	5	.029	.016	–	–	–
Eureka	31	.111	.050	–	–	–
Flow	–	–	–	530	.218	.116
Frustration	56	.309	.046	241	.101	.050
Surprise	–	–	–	56	.026	.019

5.1.2 Basic versus non-basic emotions

It is interesting to note that three out of the four emotions on our list of low frequency emotions during learning can be considered to be basic emotions; these include anger, disgust (Ekman, 2003; Izard, 1971) and contempt (Izard, 1971). This finding is consistent with other researchers who have challenged the adequacy of basing a comprehensive theory of emotions on these 'basic' emotions (Kort *et al.*, 2001; Rozin and Cohen, 2003). Although our study did not incorporate the full set of basic emotions (*i.e.*, happiness, sadness, and surprise were excluded), two recent studies by Lehman and colleagues that compared the full set of basic emotions to the learning-centred emotions and confirmed that the basic emotions were infrequent in learning sessions (Lehman *et al.*, 2008). Taken together, the results substantiate the claim that the basic emotions, although ubiquitous in everyday experience, may not be particularly relevant to learning, at least for the short learning sessions of these studies.

5.1.3 Curiosity and eureka

We suspect that curiosity might not have been experienced because students had no choice of tutoring topics in our experimental environment. If participants had been given a choice of topics, they might have picked one more relevant to their interests and displayed more curiosity. Research by Lepper and Woolverton (2002) has proposed that curiosity and engagement are systematically related to the learner's freedom of choices.

Although eureka was relatively well reported, we suspect that this response functionally signified happiness or delight from giving a correct answer rather than a deep eureka experience. True eureka experiences are more infrequent than our data suggest. In a previous study by Craig *et al.* (2004), where judges observed learners during interactions with AutoTutor, there was only one eureka experience identified in 10 h of tutoring. These concerns related to the validity of eureka was the reason why this emotion was separated into delight and surprise in the multiple judge study (Study 2).

5.1.4 Multiple-judge study

The following distribution of means emerged from the multiple judge study when averaging across affect judges (self, peer, two trained judges) and judgement types (mandatory or voluntary): Confusion, Boredom, Flow, Frustration, Delight, and Surprise, $F(5, 135) = 43.63$, $Mse = .010$, $p < .001$ (see Table 1). Once again confusion, boredom and frustration were prominent affective states that occurred during learning.

5.2 Comparison of methodologies (Goal 2)

In summary, the two studies indicate that boredom, confusion, and frustration were the most prominent affective states during interactions with AutoTutor irrespective of the affect judgement methodology (online or offline) or the affect judges (self or multiple judges). These three affective states comprised 74% of the reported emotions in the emote-aloud study and 70% of the emotions in the multiple judge study. The fact that the major affective states from the seven participants' data in the emote-aloud study was replicated in the multiple judge study with 28 participants indicates that the participants in the emote-aloud study are representative of the larger population. Although, flow was

well reported in Study 2, we cannot make any claims on its generalisability since it was not included in Study 1. As indicated earlier, requiring learners to self report on their flow experience would presumably disrupt the experience. Therefore, the subsequent analyses to determine methods for automatic detection of emotions (Goal 3) focus on boredom, confusion, and frustration, which are affective states that were frequently observed in both studies.

It is interesting to note that frustration was more prominent in the emote-aloud study than in the multiple judge study. Three possible explanations can be offered for this difference. First, the differences may be attributed to differences between the online and offline affect judgement methodologies. In online methodologies the experience of affect and verbal reports occur nearly simultaneously and are deeply grounded in the context of the learning task. The learners may be more inclined to report their frustration at the moment of the experience. This readiness to verbalise frustration can also function as a coping mechanism, *i.e.*, a way to let off some steam. In the offline affect reporting methodologies, however, learners' emotions during the judgement phase are detached from their experience. Additionally, although we attempted to simulate the contextual information with a video of the participants face and a screen capture of the session, these contextual reproductions are at best faint renditions of the actual experience.

The fact that the tutorial session for the emote-aloud study (90 min) was longer than the multiple judges (35 min) study might be the second reason why reports of frustration were more prominent for the emote-aloud study. It is reasonable to speculate that during the initial stages of the intervention participants might be willing to forge through the tutorial session, despite some frustration, with the hope that the negative emotion will be alleviated in the near future. However, if the levels of frustration are left unchecked as the session proceeds one might expect a heightening of frustration. A similar argument might be applied to the greater incidence of boredom in the emote-aloud study than in the multiple judge study. Furthermore, session length has been shown to be positively correlated with boredom (D'Mello *et al.*, 2008).

The third explanation for the lower reports of frustration in the multiple judge study may lie in the social display rules that people adhere to in expressive affect (Ekman and Friesen, 1969). Social pressures may result in the disguising of negative emotions such as frustration, thus making it difficult for judges to detect this emotion. In contrast, when encouraged to freely reflect and report on their affect, as in the emote-aloud study, such barriers drop and frustration is readily expressed.

5.3 Facial features that accompany affective expression (Goal 3)

Relationships between the action units and affective states are presented in Table 2. These patterns were extracted by performing association rule mining analyses via the *a priori* algorithm (Agarwal *et al.*, 1993) in conjunction with correlational analyses.[1] Kappa scores between the two coders for each of the AUs indicate that the level of agreement achieved by the AU judges in coding the target action units ranged from fair to excellent.[2]

The patterns between the various facial features and emotions highlight several similarities and differences between the studies. We note that the majority of the activity of the facial features during emotional experiences occurred on the upper face, with the mouth area a close second. Other facial expressions such as head nods, head shakes, and jaw movement have been excluded from Table 2 since they rarely occur.

Table 2 Patterns of facial activity accompanying affective expression

				Affective states					
		Kappa scores		*Boredom*		*Confusion*		*Frustration*	
AU	*Description*	*EL*	*MJ*	*EL*	*MJ*	*EL*	*MJ*	*EL*	*MJ*
AU1	Inner brow raiser	.94	.64					+	
AU2	Outer brow raiser	.93	.53					+	
AU4	Brow lowerer	1.00	.80			+	+		
AU7	Lid tightener	.99	.59			+	+		
AU12	Lip corner puller	.70	.71			+[a]			+[a]
AU14	Dimpler	.82	–					+[a]	
AU43	Eye closure	.77	.61	+[a]					

Notes: EL – Emote-Aloud study, MJ – Multiple Judge study, [a] Of secondary importance due to lower statistical support. + or – indicates that the AU is a positive or negative predictor of the affective state. Empty cells are indicative of no relationship between the facial feature and the affective state.

Both studies revealed similar patterns for boredom and confusion but there were differences for patterns of facial activity related to frustration. For boredom, it appears that neither study could isolate any particular subset of AUs that were associated with this emotion. In other words bored learners express no noticeable affect on their face although participants may occasionally yawn or close their eyes (Craig *et al.*, 2008). However, the incidence of yawning accompanying boredom was not frequent enough to obtain the requisite degree of statistical power to test. Yawning may also be more diagnostic of being tired than having heightened ennui.

It appears that the highly animated affective state of confusion is easily detectable from facial expressions. A lowered brow (AU4) coupled with the tightening of the lids (AU7) seems to be the prototypical expression of confusion. This pattern was replicated in both studies and so we have some confidence in the fidelity of the finding. It is tempting to speculate, from an evolutionary perspective, that learners use their face as a social cue to indicate that they are confused, in a potential effort to recruit resources from other humans to alleviate their perplexity.

In the emote-aloud study, we discovered that frustration was associated with a raised inner and outer brow (AUs 1 and 2) and a dimpler (AU 14). However, these patterns were not replicated in the multiple-judge study. This suggests that there might be occasional differences between the offline methodology employed in the multiple judge study and our previous emote-aloud methodology, which was an online measure. The emote-aloud study also included a smaller sample of participants (N = 7) when compared to the 28 participants that constituted the sample in the multiple judge study. Alternatively, the fact that the emote-aloud study utilised an online methodology where participants were encouraged to report on their emotions might explain the more visceral experiences of frustration obtained in the emote-aloud study.

5.4 Dialogue features as predictors of affect (Goal 3)

The student answer quality, tutor directness, and tutor feedback features (see above) were extracted from AutoTutor's log files and correlated with the affective states of the participants.[3] The data were selected from the turn that occurred immediately preceding or during an affective experience. The affective states were online self reports for the emote-aloud study, whereas the offline judgements were made by the self, a peer, and the two trained judges for the multiple judge study.

Table 3 indicates that systematic relationships exist between the affective states of confusion and frustration and the various dialogue features. It appears that none of our conversational measures were related to boredom.

Table 3 Patterns of dialogue features accompanying affective expression

	Affective states					
	Boredom		*Confusion*		*Frustration*	
Dialogue feature	*EL*	*MJ*	*EL*	*MJ*	*EL*	*MJ*
Student answer quality				–		+
Tutor directness			–	–		
Tutor feedback			–	–	–	–

Notes: EL – Emote-Aloud study, MJ – Multiple Judge study, + or – indicates that the dialogue feature is a positive or negative predictor of the affective state. Empty cells are indicative of no relationship between the dialogue feature and the affective state.

The tutor directness and tutor feedback feature were related to the affective state of confusion in both studies. It appears that as the feedback provided by AutoTutor leaned towards the negative direction, the learner experienced more instances of confusion. Directness of AutoTutor showed a negative relationship with confusion, so it was the tutor's hints and prompts that were affiliated with confusion rather than the tutor's assertions and summaries. Confusion is manifested much more often when the learner has to work and think. In the multiple judge study, confusion also was more readily manifested when the conceptual quality of learners' answers was not very high. This implies that confused learners were not providing very insightful answers.

As could be expected, frustration was consistently related to negative feedback being provided by the tutor. This finding was replicated across both studies and is intuitively plausible. It is interesting to note that student answer quality was positively correlated with frustration in the multiple judge study. This occurs when students have given a good answer to the immediate question, but AutoTutor's internal model of the learner's knowledge erroneously classifies the student as being a poor learner. AutoTutor responds with increased negative feedback, which in turn increases frustration (D'Mello *et al.*, 2008). This sometimes occurred when the learner had not been doing well on the topic in general and the short feedback by the tutor considered both past as well as present progress. This type of frustration could be alleviated by having AutoTutor give more positive feedback in cases where a learner takes the time to give a reasonably good response even though the learner has generally performed poorly.

6 General discussion

We have explored two methodologies to measure the emotions of a learner on the basis of human judgements, conversational cues, and facial expressions. Although there was convergence in a number of the findings across both studies, there were also some informative differences between the studies.

6.1 Advantages and disadvantages of methodologies in identifying learning-centred emotions

The emote-aloud methodology has proven to be useful for monitoring emotions while college students learn with AutoTutor. This procedure allowed us to identify the points during the AutoTutor session where affective events were occurring. The major advantage of the emote-aloud methodology is that online self reports of affect are grounded in the context of the actual affective experience.

However, there are several limitations associated with think-aloud reports of affect. First, the frequency with which affective states were reported is one potential pitfall with this methodology. Four affect states were removed from the analysis due to floor effects with reporting. These were anger, disgust, contempt, and curious. We can offer several explanations for this low incidence rate, such as the small sample size (N = 7), some hesitance to verbally report affect, and the dullness of the task (learning Computer Literacy). However, the fact remains that not all participants are amenable to the emote-aloud procedure. Some participants have a general reluctance to divulge emotional information as affective expressions are considered to be a highly socially reactive phenomenon (Bentley *et al.*, 2005).

Another limitation of the emote-aloud procedure is that it might not be sufficiently sensitive for more subtle emotions that are expressed with reduced bodily arousal. Since the affective states were reported verbally and at the learner's discretion, the verbal reports of affect generally occurred at occasions of significant physiological arousal when there was something salient to report. Therefore, one can expect physiologically charged affective states, such as confusion and frustration, to be reported more reliably for voluntary self-report measurers than less salient affective states such as boredom and flow. If this is the case, then a more standardised reporting method would be more sensitive for the less salient emotions. Such a methodology was adopted in the multiple judge study by requiring each affect judge to provide an emotion judgement every 20 sec (mandatory judgements). Additionally, by also permitting participants to simultaneously voluntarily report emotions within each 20 sec block, physiologically charged affective states can be detected as well.

Perhaps the single major concern with online verbal reports of affect is the reliance of a subjective judgement as the singular measure of the emotions of the participant. Reliance on a single operational measure of a complex construct such as emotion is problematic under even the most liberal standards for establishing construct validity. Furthermore, people may lack the requisite emotional intelligence to monitor emotions in themselves and in others (Goleman, 1997).

Some evidence to support this latter point emerged in the multiple judge study. An analysis of the interrater reliability between the self, peer, and two trained judges supports a number of conclusions about emotion measurement by humans. First, trained

judges who are experienced in coding facial actions and tutorial dialogue provide affective judgements that are more reliable ($kappa_{judge1\text{-}judge2}$ = .36) and that match the learner's self reports ($kappa_{self\text{-}judges}$ = .15) better than the judgements of untrained peers ($kappa_{self\text{-}peer}$ = .08). Second, the judgements by peers have very little correspondence to the self reports of learners. Peers apparently are not good judges of the emotions of learners (Graesser *et al.*, 2006).

Although these kappas appear to be low, the kappas for the two trained judges 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 and Forbes-Riley, 2004; Shafran *et al.*, 2003). For example, Litman and Forbes-Riley (2004) reported kappa scores of .40 in distinguishing between positive, negative, and neutral affect. Ang *et al.* (2002) reported that human judges making binary frustration-annoyance discriminations obtained a kappa score of .47. Shafran *et al.* (2003) achieved kappa scores ranging from .32 to .42 when distinguishing among six emotions. In general, these results highlight the difficulty that humans experience in detecting affect. Furthermore, it is important to understand that emotion judgements are fuzzy, ill-defined, and possibly indeterminate. A kappa score greater than 0.6 is expected when judges code some simple human behaviours, such as facial action units, basic gestures, and other visible behaviour. However, in our case the human judges are inferring a complex mental state. We argue that the lower kappa scores are meaningful, especially since it is unlikely that perfect agreement will ever be achieved and there is no objective gold standard.

In summary, the major advantage of the multiple judge study was that it supported unique perspectives with regard to the affective states of the learner. We have a higher probability of approximating the true value of the emotional construct by considering multiple models of affect. However, it should be noted that it is unclear what exactly should be the gold standard for deciding what emotions a learner is truly having. Should it be the learner or the trained judges? Although the highest interrater reliability was obtained between the trained judges, it might be nothing more than an artefact brought on by the training. Therefore, we are uncertain about the answer to this question, but it is conceivable that some emotions may best be classified by learners and others by experts. Therefore, a composite score that considers both viewpoints might be the most defensible position.

6.2 Towards automated affect detection systems

Whatever the gold standard might be, there is the challenge of developing automated affect classifiers. An automated affect classifier is of course needed to make an ITS responsive to learners' emotions. By monitoring facial features of participants during emotional expressions and dialogue features we have laid the foundation for automatic detection of the learner's affective states. Furthermore, by analysing relationships between these measures and the emotions of the learner we were able to segregate the relationships that are artefacts of the methodology (offline vs. online) from those that are true components of affect experience.

A comparison of the facial features that were diagnostic of the affective states of boredom, confusion, and frustration across both studies revealed that correspondence was discovered for boredom and confusion, but not for frustration. For boredom, the fact that none of the facial features were predictive of this state was the source of congruence across both studies. The findings for confusion were more enlightening in that the

presence of a brow lowerer and lid tightener appear to be the characteristic pattern of facial activity during episodes of confusion. This finding was first observed from the data collected with the emote-aloud protocol and was subsequently replicated in the multiple judge study. Therefore, it appears that confusion is the only affective state that can be automatically detected from the face. We have subsequently developed algorithms to segregate confusion from the baseline of neutral with an accuracy of 76% (D'Mello *et al.*, 2007). The algorithms currently rely on humans coding the Action Units but we are in the process of computing this information in an automated manner.

There also appeared to be significant relationships between the dialogue features and affective states experienced during learning. The analyses across both studies show that the directness in which speech acts are expressed by the tutor and the type of feedback given can significantly predict learners' affective states. Confusion is affiliated with indirect tutor dialogue moves (hints and prompts rather than assertions and summaries) and with negative tutor feedback. Frustration is affiliated with negative tutor feedback. It should be noted that, since negative tutor feedback is a predictor of both frustration and confusion, the tutor directness feature would be required to differentiate between these two affective states. Once again boredom did not appear to be detectable from our set of three dialogue features. Subsequently, we have engineered automated computer algorithms to discriminate confusion and frustration from a neutral baseline with accuracies of 68% and 78% respectively (D'Mello *et al.*, 2008). Similar to our results, Kapoor and colleagues report 79% accuracy in predicting frustration by monitoring a number of non-verbal and contextual channels (Kapoor *et al.*, 2007).

Of the three affective states that of primary interest to learning, boredom did not seem to be automatically detectable from facial expressions and dialogue features. However, some of our more current research is pointing towards the possibility that boredom can be detected from the learner's general body language, as tracked by a pressure sensitive chair (D'Mello *et al.*, 2007).

6.3 Concluding remarks

In this paper, we have provided an overview of the methodologies used in our research on emotions and learning with AutoTutor while reviewing three basic goals to help the literature move forward. Our investigation into the important affective states during learning yielded several consistencies emerging over our two very different methodologies. These include the consistent appearance of boredom, confusion, and frustration during the learning process and the inability to find links between the basic emotions and learning (Goal 1). However, we have also found that there can be differences based on chosen methodologies too (Goal 2). Although, we do not claim that any of the methods are superior, we have provided strengths and weakness in the hopes that these can be used as guides for choosing between these types of methodologies.

Acknowledgements

The authors gratefully acknowledge our colleagues at the University of Memphis and MIT. Special thanks to Barry Gholson, Stan Franklin, Amy Witherspoon, Jeremiah Sullins, Bethany McDaniel, Patrick Chipman, Kristy Tapp, and Brandon King, and for their valuable contributions to this study.

This research was supported by the National Science Foundation (REC 0106965, ITR 0325428, REESE 0633918). Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

References

Agarwal, R., Imielinski, T. and Swami, A. (1993) 'Mining association rules between sets of items in large databases', *Proceedings of the ACM-SIGMOID International Conference Management of Data*, Washington, DC, pp.207–216.

Ambady, N. and Rosenthal, R. (1992) 'Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis', *Psychological Bulletin*, Vol. 111, pp.256–274.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E. and Stolcke, A. (2002) 'Prosody-based automatic detection of annoyance and frustration in human-computer dialog', *Proceedings of the International Conference on Spoken Language Processing*, Denver, Colorado, pp.2037–2039.

Bentley, T., Johnston, L. and von Braggo, K. (2005) 'Evaluation using cued-recall debrief to elicit information about a user's affective experiences', *Proceedings of the OZCHI*, Canberra, Australia.

Bower, G.H. (1981) 'Mood and memory', *American Psychologist*, Vol. 36, pp.129–148.

Campbell, D.T. and Fiske, D.W. (1959) 'Convergent and discriminant validation by the multitrait-multimethod matrix', *Psychological Bulletin*, Vol. 56, pp.81–105.

Chi, M.T.H., Bassok, M., Lewis, M., Reimann, P. and Glaser, R. (1989) 'Self explanations: how students study and use examples in learning to solve problems', *Cognitive Science*, Vol. 13, pp.145–182.

Conati, C. (2002) 'Probabilistic assessment of user's emotions in educational games', *Journal of Applied Artificial Intelligence*, Vol. 16, pp.555–575.

Craig, S.D., D'Mello, S., Witherspoon, A. and Graesser, A. (2008) 'Emote-aloud during learning with AutoTutor: applying the facial action coding system to affective states during learning', *Cognition and Emotion*, Vol. 22, pp.777–788.

Craig, S.D., Graesser, A.C., Sullins, J. and Gholson, B. (2004) 'Affect and learning: an exploratory look into the role of affect in learning', *Journal of Educational Media*, Vol. 29, pp.241–250.

Csikszentmihalyi, M. (1990) *Flow: The Psychology of Optimal Experience*, New York: Harper-Row.

D'Mello, S.K., Craig, S.D., Gholson, B., Franklin, S., Picard, R. and Graesser, A.C. (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.

D'Mello, S.K., Craig, S.D., Sullins, J. and Graesser, A.C. (2006) 'Predicting affective states through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue', *International Journal of Artificial Intelligence in Education*, Vol. 16, pp.3–28.

D'Mello, S.K., Craig, S.D., Witherspoon, A.W., McDaniel, B.T. and Graesser, A.C. (2008) 'Automatic detection of learner's affect from conversational cues', *User Modeling and User-Adapted Interaction*, Vol. 18, Nos. 1–2, pp.45–80.

D'Mello, S.K., Picard, R. and Graesser, A.C. (2007) 'Towards an affect sensitive AutoTutor', *IEEE Intelligent Systems*, Vol. 22, No. 4, pp.53–61.

De Vicente, A. and Pain, H. (2002) 'Informing the detection of students' motivational state: an empirical study', *Intelligent Tutoring Systems 2002*, Berlin, Germany, pp.933–943.

Ekman, P. (2003) *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*, New York: Henry Holt and Company, LLC.

Ekman, P. and Friesen, W.V. (1969) 'The repertoire of nonverbal behavior. Categories, origins, usage, and coding', *Semiotica*, Vol. 1, pp.49–98.

Ekman, P. and Friesen, W.V. (1978) *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Palo Alto: Consulting Psychologists Press.

Ericsson, K.A. and Simon, H.A. (1993) *Protocol Analysis: Verbal Reports as Data*, Revised edition, Cambridge, MA: The MIT Press.

Eva-Wood, A.L. (2004) 'How think-and feel-aloud instruction influences poetry readers', *Discourse Processes*, Vol. 38, pp.173–192.

Fredrickson, B.L. and Branigan, C. (2005) 'Positive emotions broaden the scope of attention and thought-action repertoires', *Cognition and Emotion*, Vol. 19, pp.313–332.

Gee, J.P. (2003) *What Video Games Have to Teach Us About Language and Literacy*, New York: Macmillan.

Goleman, D. (1997) *Emotional Intelligence*, NewYork: Bantam Books.

Graesser, A.C., Chipman, P., King, B., McDaniel, B. and D'Mello, S. (2007a) 'Emotions and learning with AutoTutor', in R. Luckin *et al.* (Eds.) *13th International Conference on Artificial Intelligence in Education (AIED 2007)*, IOS Press, pp.569–571.

Graesser, A.C., Jackson, G.T. and McDaniel, B. (2007b) 'AutoTutor holds conversations with learners that are responsive to their cognitive and emotional states', *Educational Technology*, Vol. 47, pp.19–22.

Graesser, A.C., Lu, S., Jackson, G.T., Mitchell, H., Ventura, M., Olney, A. and Louwerse, M.M. (2004) 'AutoTutor: a tutor with dialogue in natural language', *Behavioral Research Methods, Instruments, and Computers*, Vol. 36, pp.180–193.

Graesser, A.C., Lu, S., Olde, B.A., Cooper-Pye, E. and Whitten, S. (2005) 'Question asking and eye tracking during cognitive disequilibrium: comprehending illustrated texts on devices when the devices break down', *Memory and Cognition*, Vol. 33, pp.1235–1247.

Graesser, A.C., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S. and Gholson, B. (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.

Graesser, A.C., Person, N., Harter, D. and the Tutoring Research Group (2001) 'Teaching tactics and dialogue in AutoTutor', *International Journal of Artificial Intelligence in Education*, Vol. 12, pp.257–279.

Grimm, M., Mower, E., Kroschel, K. and Narayan, S. (2006) 'Combining categorical and primitives-based emotion recognition', *14th European Signal Processing Conference (EUSIPCO)*, Florence, Italy.

Izard, C.E. (1971) *The Face of Emotion*, New York: Appleton-Century-Crofts.

Kapoor, A., Burleson, W. and Picard, R.W. (2007) 'Automatic prediction of frustration', *International Journal of Human-Computer Studies*, Vol. 65, pp.724–736.

Klein, J., Moon, Y. and Picard, R. (2002) 'This computer responds to user frustration – theory, design, and results', *Interacting with Computers*, Vol. 14, pp.119–140.

Kort, B., Reilly, R. and Picard, R. (2001) 'An affective model of interplay between emotions and learning: reengineering educational pedagogy – building a learning companion', in T. Okamoto, R. Hartley, R.H. Kinshuk and J.P. Klus (Eds.) *Proceedings IEEE International Conference on Advanced Learning Technology: Issues, Achievements and Challenges*, Madison, Wisconsin: IEEE Computer Society, pp.43–48.

Landauer, T.K. and Dumais, S.T. (1997) 'A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge', *Psychological Review*, Vol. 104, pp.211–240.

Larsen, J.T., McGraw, A.P. and Cacioppo, J.T. (2001) 'Can people feel happy and sad at the same time?', *Journal of Personality and Social Psychology*, Vol. 81, pp.684–696.

Lehman, B.A., Matthews, M., D'Mello, S.K. and Person, N. (2008) 'Understanding students' affective states during learning', in B. Woolf *et al.* (Eds.) *Ninth International Conference on Intelligent Tutoring Systems*, ITS 2008, LNCS 5091, Springer-Verlag, pp.50–59.

Lepper, M.R. and Chabay, R.W. (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.

Lepper, M.R. and Henderlong, J. (2000) 'Turning "play" into "work" and "work" into "play": 25 years of research on intrinsic versus extrinsic motivation', in C. Sansone and J.M. Harackiewicz (Eds.) *Intrinsic and Extrinsic Motivation: The Search for Optimal Motivation and Performance*, San Diego, CA: Academic Press, pp.257–307.

Lepper, M.R. and Woolverton, M. (2002) 'The wisdom of practice: Lessons learned from the study of highly effective tutors', in J. Aronson (Ed.) *Improving Academic Achievement: Impact of Psychological Factors on Education*, Orlando, FL: Academic Press, pp.135–158.

Linnenbrink, E.A. and Pintrich, P.R. (2002) 'The role of motivational beliefs in conceptual change', in M. Limon and L. Mason (Eds.) *Reconsidering Conceptual Change: Issues in Theory and Practice*, Dordretch, the Netherlands: Kluwer Academic Publishers, pp.115–135.

Litman, D.J. and Forbes-Riley, K. (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.

Mandler, G. (1999) 'Emotion', in B.M. Bly and D.E. Rumelhart (Eds.) *Cognitive Science. Handbook of Perception and Cognition*, 2nd ed., San Diego, CA: Academic Press, pp.367–384.

McDaniel, B.T., D'Mello, S.K., King, B.G., Chipman, P., Tapp, K. and Graesser, A.C. (2007) 'Facial features for affective state detection in learning environments', in D.S. McNamara and J.G. Trafton (Eds.) *Proceedings of the 29th Annual Cognitive Science Society*, Austin, TX: Cognitive Science Society, pp.467–472.

McQuiggan, S.W. and Lester, J.C. (2007) 'Leveraging affect for narrative-centered guided discovery learning environments', *Supplementary Proceedings of the 13th International Conference on Artificial Intelligence in Education (AIED 2007)*, pp.67–74.

Meyer, D.K. and Turner, J.C. (2006) 'Reconceptualizing emotion and motivation to learn in classroom contexts', *Educational Psychology Review*, Vol. 18, pp.377–390.

Miserandino, M. (1996) 'Children who do well in school: individual differences in perceived competence and autonomy in above-average children', *Journal of Educational Psychology*, Vol. 88, pp.203–214.

Mota, S. and Picard, R.W. (2003) 'Automated posture analysis for detecting learner's interest level', *Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction, CVPR HCI*, June.

Newell, A. and Simon, H.A. (1972) *Human Problem Solving*, Englewood Cliffs, NJ: Prentice Hall.

Norman, D.A. (1994) 'How might people interact with agents?', *Communication of the ACM*, Vol. 37, No. 7, pp.68–71.

Ortony, A., Clore, G.L. and Collins, A. (1988) *The Cognitive Structure of Emotions*, New York: Cambridge University Press.

Paiva, A., Prada, R. and Picard, R.W. (Eds.) (2007) *Affective Computing and Intelligent Interaction*, Springer.

Pantic, M. and Rothkrantz, L.J.M. (2003) 'Towards an affect-sensitive multimodal human-computer interaction', *IEEE, Special Issue on Multimodal Human-Computer Interaction (HCI)*, Vol. 91, No. 9, pp.1370–1390.

Patrick, B., Skinner, E. and Connell, J. (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*, Vol. 65, pp.781–791.

Pekrun, R., Goetz, T., Titz, W. and Perry, R.P. (2002) 'Academic emotions in students' self-regulated learning and achievement: a program of qualitative and quantitative research', *Educational Psychologist*, Vol. 37, No. 2, pp.91–105.

Picard, R.W. (1997) *Affective Computing*, Cambridge, MA: MIT Press.

Robson C. (1993) *Real Word Research: A Resource for Social Scientist and Practitioner Researchers*, Oxford: Blackwell.

Rozin, P. and Cohen, A.B. (2003) 'High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans', *Emotion*, Vol. 3, pp.68–75.

Russell, J.A. (2003) 'Core affect and the psychological construction of emotion', *Psychological Review*, Vol. 110, pp.145–172.

Shafran, I., Riley, M. and Mohri, M. (2003) 'Voice signatures', *Proceedings IEEE Automatic Speech Recognition and Understanding Workshop*, Piscataway, NJ: IEEE, pp.31–36.

Shneiderman, B. and Plaisant, C. (2005) *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 4th ed., Reading, MA: Addison-Wesley.

Silvia, P. and Abele, A. (2002) 'Can positive affect induce self-focused attention? Methodological and measurement issues', *Cognition and Emotion*, Vol. 16, pp.845–853.

Stein, N.L. and Hernandez, M.W. (in press) 'Assessing understanding and appraisals during emotional experience: the development and use of the Narcoder', in J.A. Coan and J.J. Allen (Eds.) *Handbook of Emotion Elicitation and Assessment*, New York: Oxford University Press.

Storey, J.K., Kopp, K.J., Wiemer, K., Chipman, P. and Graesser, A.C. (in press) 'Critical thinking tutor: tsing AutoTutor to teach scientific critical thinking skills', *Behavioral Research Methods*.

Strahan, C. and Zytowski, D.G. (1976) 'Impact of visual, vocal, and lexical cues on judgments of counselor qualities', *Journal of Counseling Psychology*, Vol. 23, pp.387–393.

Thayer, R.E. (1989) *The Biopsychology of Mood and Activation*, New York: Oxford University Press.

Trabasso, T. and Magliano, J. (1996) 'Conscious understanding during comprehension', *Discourse Processes*, Vol. 21, pp.225–286.

VanLehn, K., Graesser, A.C., Jackson, G.T., Jordan, P., Olney, A. and Rose, C.P. (2007) 'When are tutorial dialogues more effective than reading?', *Cognitive Science*, Vol. 30, pp.3–62.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T. and Baggett, W.B. (2003) 'Why do only some events cause learning during human tutoring?', *Cognition and Instruction*, Vol. 21, No. 3, pp.209–249.

Vorderer, P. (2003) 'Entertainment theory', in B. Jennings, D. Roskos-Ewoldsen and J. Cantor (Eds.) *Communication and Emotion: Essays in Honor of Dolf Zillmann*, Mahwah, NJ: Erlbaum, pp.131–153.

Woolf, B., Burelson, W. and Arroyo, I. (2007) 'Emotional intelligence for computer tutors', in S. D'Mello, S.D. Craig, R. El Kaliouby, M. Alsmeyer and G. Rebolledo-Mendes (Eds.) *Workshop on Modeling and Scaffolding Affective Experiences to Impact Learning*, pp.6–15, http://www.informatics.sussex.ac.uk/users/gr20/aied07/AffectWkshpAIED07Proceedings-R1.pdf (retrieved 3 July 2007).

Notes

1 Additional details on the FACS coding process for the emote-aloud study is reported in Craig *et al.*, 2008). For the multiple judge study McDaniel *et al.* (2007) provide an in depth discussion on the coding procedure and statistical analyses.

2 The Kappa statistic measures the proportion of agreement between two raters with correction for chance. 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, 1993).

3 Additional details on the dialogue features and statsistical analyses for the emote-aloud data are reported in D'Mello *et al.* (2006). For the multiple judge study analysis details appear in D'Mello *et al.* (2008).

Appendix

Definitions of emotions used in studies

Affective state	*Definition*
Anger[1]	Strong feeling of displeasure and usually of antagonism
Boredom[1,2]	State of being weary and restless through lack of interest
Confusion[1,2]	Failure to differentiate similar or related ideas/ noticeable lack of understanding
Contempt[1]	The act of despising, a lack of respect or reverence for something
Curious[1]	An active desire to learn or to know
Disgust[1]	Marked aversion aroused by something highly distasteful
Eureka[1]	A feeling used to express triumph on a discovery
Flow[2]	State of interest that results from involvement in an activity
Frustration[1,2]	Making vain or ineffectual efforts however vigorous; a deep chronic sense or state of insecurity and dissatisfaction arising from unresolved problems or unfulfilled needs; dissatisfaction or annoyance
Neutral[2]	No apparent emotion or feeling
Surprise[2]	Wonder or amazement, especially from the unexpected

Notes: [1] Used in Study 1, [2] Used in Study 2.