Multi-Method Assessment of Affective Experience and Expression during Deep Learning

Sidney K. D’Mello, Scotty D. Craig, and Art C. Graesser

University of Memphis

Keywords. affect, affective states, emotions, learning, dialogue features, facial features, confusion, frustration, boredom, online, offline, emote-aloud
Abstract

Inquiries into the link between affect and learning require robust methodologies to measure the learner’s affective states. We describe two studies that utilized either an online or an 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 judgments while the offline study considered judgments by the learner, a peer, and two trained judges. The studies also investigated relationships between facial features, conversational cues and emotional expressions in an attempt to scaffold the development of computer algorithms to automatically detect the 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.
Multi-Method Assessment of Affective Experience and Expression during Deep Learning

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 memorizing definitions, facts, and properties (i.e. shallow learning) (Graesser, Jackson, & McDaniel, 2007). 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 & Henderlong, 2000; Linnenbrink & Pintrich, 2004; Meyer & Turner, 2006; Stein & 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, Picard, & Graesser, 2007; Graesser, McDaniel, & Jackson, 2007; Kort, Reilly, & Picard, 2001; Lepper & 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, Reilly, & Picard, 2001; Litman & Forbes-Riley, 2004; McQuiggan & Lester 2007; Woolf, Burelson, & Arroyo, 2007). For example, Kort, Reilly, and Picard (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 recognizes a learner’s affect by monitoring facial features, posture
patterns, and onscreen keyboard/mouse behaviors. 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 colleagues work with their ITSPROE (2004) 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; D’Mello, Picard, & Graesser, 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 artifacts that range from affect-sensitive ITSs on technical material (D’Mello et al., 2005; 2007; Kort, Reilly, & Picard, 2001; Litman & Forbes-Riley, 2004; Woolf, Burelson, & Arroyo, 2007) to entertaining media and serious games (Conati, 2002; Gee, 2003; McQuiggan & Lester, 2007; Vorderer et al., 2004). 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: (1) identifying some of the important states that occur during learning, (2) comparing methodologies for monitoring these affective states, and (3) developing systems to automatically detect the affective states in real time.
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, Clore, and Collins (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 & Pintrich, 2004). Researchers in different fields are familiar with Ekman’s pioneering work on the detection of emotions from facial expressions (Ekman & 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 hour learning sessions (D’Mello, Craig, Witherspoon, McDaniel, & Graesser, 2006; Graesser et al., 2007; Lehman, Matthews, D’Mello, & Person, 2008; Pekrun, Goetz, Titz, & Perry 2002). Researchers have proposed a different set of emotions that prevail during complex learning, namely boredom (Miserandino, 1996), confusion (Craig, Graesser, Sullins, & Gholson, 2004; Graesser, Lu et al., 2005; Kort, Reilly, & Picard, 2001), frustration (Kort et al., 2001; Patrick et al, 1993), delight (Fredrickson & Branigan, 2005; Silvia & Abele, 2002), and flow (Csikszentmihalyi, 1990). Empirical tests of this list of learning-centered affective states is the first goal of our research.

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 judgment (self reports or external observers) and the time (online or offline) of the judgment. 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, McGraw, & Cacioppo, 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 judgments 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 judgments 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, Lu, Olde, Cooper-Pye, & Whitten, 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 re-inspecting the
behaviors at a later time. Additionally, there were no self reports or additional measures that the observed judgments 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.

**Goal 3. Automated Detection of Learning-Centered 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 recognize 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 & Plaisant, 2005). Therefore, 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 endeavor because emotional expressions are sometimes murky, subtle, and compounded with individual differences in experience and expression. Therefore, robust recognition of the 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 & Rothkrantz, 2003 for a comprehensive review and Paiva, Prada, & Picard, 2007 for recent updates). Most of these affect-detection systems attempt to
recognize 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-centered 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 artifact 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.

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’ verbalized their emotions while interacting with the tutor (D’Mello, Craig, Sullins, & Graesser, 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 judgment 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 versus 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-centered 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 generalize to Study 2. This can be accomplished by analyzing each study independently, identifying reliable patterns, and assessing whether these patterns are observed across studies. Any patterns that do generalize can be attributed to an inherent characteristic of affect-learning interactions and not an artifact 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 organized 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 generalizable 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-centered affective states.

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, Jackson, & McDaniel, 2005; Graesser, Person, Harter, & TRG, 2001). AutoTutor attempts to comprehend the students’ natural language contributions and then responds to the 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 summarizes 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, Lu et al., 2004), conceptual physics (VanLehn, Graesser, et al., 2007), and critical reasoning on scientific methods (Storey, Kopp, Wiemer, Chipman, & Graesser, 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, Lu et al., 2004; VanLehn, Graesser, et al., 2007). Moreover, the goals of the present study were to analyze 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.

**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, D’Mello, Witherspoon, & Graesser, 2008; D’Mello, Craig, Sullins, & Graesser, 2006). Although, this method is based on the subjective report of the learner, we hypothesize 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 & 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 & Simon, 1993), comprehending text (Trabasso & 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 on-line 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 labor-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 generalize 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).

**Brief Sketch of Methodology**

**Measuring Affective States.** The emote-aloud study had 7 learners who were video taped while interacting with the AutoTutor system for approximately 1.5 hours. 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.

Coding Facial Expressions. Ekman and Friesen’s (1978) Facial Action Coding System (FACS) was adopted when we analyzed the facial features that accompanied the various emotions. FACS specifies how judges are to code specific facial behaviors (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 3 seconds (Ekman & Friesen, 1978). Therefore, two coders independently scored the three seconds before an emote-aloud report was made using FACS (see Craig, D’Mello, Witherspoon, & Graesser, 2008 for more details).

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 realized by 3-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 & Dumais, 1997) and other semantic components. The answer quality measure was the semantic match between the learners’ response and the expected, ideal response from the student. 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 students’ 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.

**Offline Emotion Judgments 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 & Pain, 2002; Klein, Moon, & Picard, 2002) or ratings by independent judges (Litman & Forbes-Riley, 2004; Mota & 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 & 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 judgments occurred after the interaction session. Finally, neutral was added because participants were required to make forced-choice affect judgments.

**Brief Sketch of Methodology**

Study 2 had 28 college students who interacted with the AutoTutor system for 35 minutes. A video of the learners’ faces and a video of their computer screen were recorded. The judging process was initiated by synchronizing the video streams from the screen and the face, and displaying them to the judges. Judges were instructed to make judgments on what affective states were present in 20-second intervals (*mandatory* judgments), at which time the video
automatically paused. They were also instructed to indicate any affective states that were present in between the 20-second stops (voluntary judgments). Mandatory and voluntary judgments occurred within the same judging session and judges were able to replay the 20 second segment prior to making an emotion judgment. 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 seconds long (Strahan & Zytowski, 1976). Our consistent sampling rate of 20 seconds was sufficiently long to identify the learner’s affect and also allowed ample time for valid voluntary judgments to occur.

Four sets of emotion judgments were made for each learner’s AutoTutor session. For the self judgments, learners watched their own sessions with AutoTutor immediately after having interacted with the tutor. For the peer judgments, 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 analyzed 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 & 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 judgments.

Coding Facial Expressions and Dialogue Features. The process of coding of facial and dialogue features was the same as the emote-aloud study.
Results and Discussion

The Incidence of Affective States (Goal 1)

*Emote-Aloud Study.* In the emote-aloud study, there was a significant difference in the proportions of the various 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.

*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, Reilly, & Picard, 2001; Rozin & 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-centered emotions and confirmed that the basic emotions were infrequent in learning sessions (Lehman, Matthews, D’Mello, & Person, 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.
Table 1

*Proportion of Affective States Observed*

<table>
<thead>
<tr>
<th>Affective States</th>
<th>Emote-Aloud Study (Online)</th>
<th>Multiple Judge Study (Offline)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>Proportions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>Anger</td>
<td>17</td>
<td>.054</td>
</tr>
<tr>
<td>Boredom</td>
<td>43</td>
<td>.260</td>
</tr>
<tr>
<td>Confusion</td>
<td>54</td>
<td>.171</td>
</tr>
<tr>
<td>Contempt</td>
<td>8</td>
<td>.061</td>
</tr>
<tr>
<td>Curious</td>
<td>1</td>
<td>.003</td>
</tr>
<tr>
<td>Delight</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Disgust</td>
<td>5</td>
<td>.029</td>
</tr>
<tr>
<td>Eureka</td>
<td>31</td>
<td>.111</td>
</tr>
<tr>
<td>Flow</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Frustration</td>
<td>56</td>
<td>.309</td>
</tr>
<tr>
<td>Surprise</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*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 hours 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).

Multiple-Judge Study. The following distribution of means emerged from the multiple judge study when averaging across affect judges (self, peer, 2 trained judges) and judgment 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.

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 judgment 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 7 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 generalizability 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 judgment 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 verbalize frustration can also function as a coping mechanism, i.e. a way to let off some steam. In the offline affect reporting methodologies, however, the learners’ emotions during the judgment phase are detached from the 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 minutes) was longer than the multiple judges (35 mins) 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, Craig, Witherspoon, McDaniel, & Graesser, 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 & 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.

**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 & Srikant, 1994) 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.

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, D’Mello, Witherspoon, & Graesser, 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.

Table 2

*Patterns of Facial Activity Accompanying Affective Expression.*

<table>
<thead>
<tr>
<th>AU</th>
<th>Description</th>
<th>Kappa Scores</th>
<th>Affective States</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EL</td>
<td>MJ</td>
</tr>
<tr>
<td>AU1</td>
<td>Inner Brow Raiser</td>
<td>.94</td>
<td>.64</td>
</tr>
<tr>
<td>AU2</td>
<td>Outer Brow Raiser</td>
<td>.93</td>
<td>.53</td>
</tr>
<tr>
<td>AU4</td>
<td>Brow Lowerer</td>
<td>1.00</td>
<td>.80</td>
</tr>
<tr>
<td>AU7</td>
<td>Lid Tightener</td>
<td>.99</td>
<td>.59</td>
</tr>
<tr>
<td>AU12</td>
<td>Lip Corner Puller</td>
<td>.70</td>
<td>.71</td>
</tr>
<tr>
<td>AU14</td>
<td>Dimpler</td>
<td>.82</td>
<td>-</td>
</tr>
<tr>
<td>AU43</td>
<td>Eye closure</td>
<td>.77</td>
<td>.61</td>
</tr>
</tbody>
</table>

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.

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 on-line 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 utilized 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.

*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 judgments were made by the self, a peer, and the 2 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*

<table>
<thead>
<tr>
<th>Dialogue Feature</th>
<th>Boredom</th>
<th>Confusion</th>
<th>Frustration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EL</td>
<td>MJ</td>
<td>EL</td>
</tr>
<tr>
<td>Student Answer Quality</td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Tutor Directness</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Tutor Feedback</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*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 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 learner’s 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.

**General Discussion**

We have explored two methodologies to measure the emotions of a learner on the basis of human judgments, 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.

*Advantages and Disadvantages of Methodologies in Identifying Learning-Centered 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 standardized 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 judgment every 20 seconds (mandatory judgments). Additionally, by also permitting participants to simultaneously voluntarily report emotions within each 20 second 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 judgment 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. 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 2 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 judgments that are more reliable ($\kappa_{\text{judge1-judge2}} = .36$) and that match the learner’s self reports ($\kappa_{\text{self-judges}} = .15$) better than the judgments of untrained peers ($\kappa_{\text{self-peer}} = .08$). Second, the judgments 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 2 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 & Forbes-Riley, 2004; Shafran, Riley, & Mohri, 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, Riley, and Mohri (2003) achieved kappa scores ranging from .32 to .42 when distinguishing among 6 emotions. In general, these results highlight the difficulty that humans experience in detecting affect. Furthermore, it is important to understand that emotion judgments are fuzzy, ill-defined, and possibly indeterminate. A kappa score greater than 0.6 is expected when judges code some simple human behaviors, such as facial action units, basic gestures, and other visible behavior. 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 artifact 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.

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 analyzing relationships between these measures and the emotions of the learner we were able to segregate the relationships that are artifacts 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, Picard, & Graesser, 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 the 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, Burleson, & Picard, 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 toward the possibility that boredom can be detected from the
learner’s general body language, as detected by a pressure sensitive chair (D’Mello, Picard, & Graesser, 2007).

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.
References


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


http://www.informatics.sussex.ac.uk/users/gr20/aied07/AffectWkshpAIED07Proceedings-R1.pdf
Author Note

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.

Correspondence concerning this article should be addressed to Sidney D’Mello, Department of Computer Science, University of Memphis, Memphis, TN 38152, USA. Email: sdmello@memphis.edu.
Appendix

Definitions of Emotions Used in Studies

<table>
<thead>
<tr>
<th>Affective State</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger(^1)</td>
<td>strong feeling of displeasure and usually of antagonism</td>
</tr>
<tr>
<td>Boredom(^{1,2})</td>
<td>state of being weary and restless through lack of interest.</td>
</tr>
<tr>
<td>Confusion(^{1,2})</td>
<td>failure to differentiate similar or related ideas/ noticeable lack of understanding</td>
</tr>
<tr>
<td>Contempt(^1)</td>
<td>the act of despising, a lack of respect or reverence for something</td>
</tr>
<tr>
<td>Curious(^1)</td>
<td>an active desire to learn or to know</td>
</tr>
<tr>
<td>Disgust(^1)</td>
<td>marked aversion aroused by something highly distasteful</td>
</tr>
<tr>
<td>Eureka(^1)</td>
<td>a feeling used to express triumph on a discovery</td>
</tr>
<tr>
<td>Flow(^2)</td>
<td>state of interest that results from involvement in an activity</td>
</tr>
<tr>
<td>Frustration(^{1,2})</td>
<td>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</td>
</tr>
<tr>
<td>Neutral(^2)</td>
<td>no apparent emotion or feeling</td>
</tr>
<tr>
<td>Surprise(^2)</td>
<td>wonder or amazement, especially from the unexpected</td>
</tr>
</tbody>
</table>

Notes: \(^1\) Used in Study 1, \(^2\) Used in Study 2
Footnotes

1 Additional details on the FACS coding process for the emote-aloud study is reported in Craig, D’Mello, Witherspoon, & Graesser, 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, 2003).

3 Additional details on the dialogue features and statistical analyses for the emote-aloud data are reported in D’Mello, Craig, Sullins, & Graesser, 2006. For the multiple judge study analysis details appear in D’Mello et al., 2008.