Emote-Aloud during Learning with AutoTutor:
Applying the Facial Action Coding System to Cognitive-Affective States during Learning
Scotty D. Craig, Sidney D’Mello, Amy Witherspoon, & Art Graesser

Contact person:
Art Graesser
202 Psychology building
The University of Memphis
Memphis, TN 38152
Phone: 901 678-2742
Fax: 901 678-2579
a-graesser@memphis.edu
Abstract

In an attempt to discover the facial action units for affective states that occur during complex learning, this study adopted an emote-aloud procedure in which participants were recorded as they verbalized their affective states while interacting with an intelligent tutoring system (AutoTutor). Participants’ facial expressions were coded by two expert raters using Ekman’s Facial Action Coding System and analyzed using association rule mining techniques. The two expert raters received an overall Kappa that ranged between .76 to .84. The association rule mining analysis uncovered facial actions associated with confusion, frustration, and boredom. We discuss these rules and the prospects of enhancing AutoTutor with non-intrusive affect-sensitive capabilities.
Automated detection of affective states is on the horizon of computer science and engineering, in subfields that range from intelligent tutoring systems that help people learn to security systems that attempt to detect terrorists. In the arena of learning, there are sensing systems with computational algorithms that can classify the learners’ affective states as they interact with tutoring systems (Fan et al., 2003) and educational games (Conati, 2002). The accuracy of these systems is modest, but improving.

Detection of a learner’s affective states has proven to be difficult task because there are many channels of communication (i.e., dialogue, facial expressions, speech) and a learner’s reaction to learning material can change as a function of their goals, preferences, expectations, exposure, and knowledge states. Attentive human tutors (Lepper, Woolverton, Mumme, & Gurtner, 1991) and teachers (Meyer & Turner, 2002) are presumably able to detect these changing affective states and to modify their pedagogical tactics. In the case of automated learning environments, computer systems will need to detect different affective states that are important during learning in order to respond effectively to the learners.

The present study investigated two methods of identifying affective states while learners interacted with AutoTutor. AutoTutor (Graesser, Chipman, Haynes, & Olney, 2005) is an automated computer tutor that simulates human tutors and holds conversations with students in natural language. It helps students learn by presenting problems (or questions) from a curriculum script and engaging in a mixed-initiative dialogue during the construction of an answer by using assertions, answers to student questions, corrections of student errors,
hints, pumps (e.g., tell me more), prompts to get the student to fill in missing words, and summaries. One method of determining learners’ affective states consists of self-reports by the learner, a method called *emote aloud*. College students were recorded as they verbalized their affective states during learning with AutoTutor. Our second method had trained judges analyze the facial expressions during learning with Ekman’s Facial Action Coding System (Ekman, 2003; Ekman & Friesen, 1978). We then examined how the facial actions were systematically correlated with the affective states manifested in the emote-aloud procedure.

Ekman and Friesen’s (1978) Facial Action Coding System specifies how coding specific facial behaviors, based on the muscles that produce them, could help identify the “basic emotions” of happiness, sadness, surprise, disgust, anger, and fear (Ekman & Friesen, 1978). Each movement in the face is called an *action unit* (or AU) and there is a total of 58 AUs. Some limitations to the coding system arise from the fact that it was tested on static pictures rather than on changing expressions over time. Some researchers have also challenged the adequacy of basing a complete theory of emotions on these “basic” emotions (Rozin & Cohen, 2003). Moreover, these six emotions are not the most frequent nor significant emotions in the learning process (Craig, Graesser, Sullins, & Gholson, 2004; Kapoor, Mota, & Picard, 2001; Pekrun, Goetz, Titz, & Perry, 2002).

Researchers are only beginning to identify the prominent affective states that occur during complex learning. Linnenbrink and Pintrich (2004), for example, investigated the link between general affect measures (positive and negative) and learning while college students read passages on Newtonian physics. They found no relationship between reports of positive affect and learning, but learners’ physics understanding scores decreased as a function of negative affect during learning. Craig et al. (2004) documented some of the affective states
that occur in college students while learning the fundamentals of computers literacy with AutoTutor. Trained judges observed whether participants were experiencing frustration, boredom, flow/interest, confusion, eureka, and neutral during interactions with AutoTutor. Correlations between the frequency of each affect state and learning gains revealed that learning gains were positively correlated with flow/engagement and confusion, but negatively correlated with boredom. These studies are hardly the final word on what affective states occur during learning and what relations exist between these states and learning gains.

The present study focused on the affective states of anger, boredom, confusion, contempt, curiosity, disgust, eureka, and frustration. These affect states might be viewed cognitive states rather than pure affective states, but some researchers have classified them as affect states (Barrett, 2006; Meyer & Turner, 2002; Stein, & Hernandez, in press). Our position agrees with these latter researchers because these states are accompanied by enhanced physiological arousal compared with a neutral base rate state (Damasio, 1995; LeDoux, 1996) and because affect-cognition blends are relevant to complex learning (Ortony, Clore, & Collins, 1988; Isen, 1999). The selection of many of these affective states was motivated by the observational study by Craig et al. 2004) in which trained judges observed affective states that learners appeared to experience while interacting with AutoTutor. However, there also are theoretical motivations for selecting these states. For example, curiosity has a long history of importance to learning in the motivation literature (Stipek, 1998), whereas the elevated positive affective state of eureka is experienced during moments of insight during problem solving (Knoblich, Ohlsson, & Raney, 2001). Confusion and sometimes frustration occur when individuals experience cognitive disequilibrium from obstacles to goals, breakdowns in systems, contradictions, anomalies, and salient gaps in
knowledge (Graesser, Lu, Olde, Cooper-Pye, & Whitten, 2005; Otero & Graesser, 2001). Frustration from being stuck is often investigated in the automated affect-detection literature (Litman & Forbes-Riley, 2006; Picard, 1997). The extreme negative affective states of anger, contempt and disgust were selected because they were viewed as increasing degrees of frustration (Ekman, 2003) that could inhibit learning in view of the Linnenbrink and Pintrich (2004) results on affective states while learning physics (see also, Pekrun et al., 2002). These negative emotions are also aligned with the negative basic emotions (Ekman & Friesen, 1978).

Our *emote-aloud* procedure is a modification of a think-aloud procedure that is frequently used in cognitive science (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), reading poetry (Eva-Wood, 2004), reading English literature (Earthman, 1992) or solving break down scenarios on electronic equipment (Graesser et al., 2005). 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 affective states while working on a task, with minimal task interference.

Think-aloud studies and this current 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. Moreover, the present study collected and analyzed the facial actions of learners, an equally labor-intensive process. The number of participants can be small in the methodology using the action unit coding, yet still yield rich and reliable data (Ekman, 2003).

Methods

Participants

The participants were 7 undergraduates in the department of psychology subject pool at the University of Memphis. Two participants were discarded because they expressed so few affective states that their contributions yielded insufficient data to investigate our main questions. However, this does raise the important empirical point that not all participants are amenable to the emote-aloud procedure.

Materials

*Electronic materials.* Participants interacted with AutoTutor (Graesser, et al., 2005) on topics in computer literacy. This conversation with AutoTutor typically took between 30-100 conversational turns.

*Emote-aloud procedure.* The participants were video recorded while interacting with AutoTutor and providing the emote-aloud protocols. They were instructed to say aloud the affective state whenever they experienced an affective state. Participants were given eight affective states (anger, boredom, confusion, contempt, curiosity, disgust, eureka, and frustration), which were functionally defined for the participants. Anger was defined as a strong feeling of displeasure and usually of antagonism. Boredom was defined as the state of being weary and restless through lack of interest. Confusion was defined as the failure to differentiate from an often similar or related other. Contempt was defined as the act of
despising, a lack of respect or reverence for something. Curious was defined as an active desire to learn or to know. Disgust was defined as marked aversion aroused by something highly distasteful. Eureka was defined as a feeling used to express triumph on a discovery. Frustration was defined as making vain or ineffectual efforts however vigorous; a deep chronic sense or state of insecurity and dissatisfaction arising from unresolved problems or unfulfilled needs. All definitions were taken from Merriam-Webster online (2003).

Attempts were made to keep the emote-aloud procedure as natural as possible. Participants were allowed to verbalize their affective states (either single states or in extremely rare cases multiple states) without external prompts from the system or pressure from an experimenter in the room. The experimenter observed via a one way mirror and could check up on the participant if long pauses were observed.

Procedure

As participants came into the lab, they were instructed they would be expressing their affective states while learning about computer literacy with a computer system called AutoTutor. The participants were given the list of affective states along with definitions provided above. Participants were asked to interact normally with the AutoTutor system and to emote whenever they experienced one of the affective states. The participants then interacted with AutoTutor for 90 minutes and engaged in the emote-aloud activity.

Data Treatment

Scoring procedure. Emote-aloud episodes were identified in each participant’s video by extracting 10-second clips that ended in an emote-aloud utterance. No obvious ambiguities were encountered in the participants’ articulation of the targeted affective states because the affective states had been defined prior to the tutoring session. The participants’
self reports of affect states were assumed to be valid. The expression of affective states tended to be a short duration, rarely more than 3 seconds (Ekman, 2003). Two judges independently scored the three seconds before the utterance was made using the Facial Action Coding System (Ekman & Friesen, 1978). That is, for the three seconds before an utterance was made, the judges watched the clips and recorded AUs and the time of observation. Clips with multiple affective states (N = 4) were not included in the analyses. This resulted in a total of 201 valid clips.

An AU database was created from the judges’ coding of AUs. Each record in the database consisted of one or more AU from the same clip based on the time stamp in which they were observed. This allowed for multiple AU records for the same emote-aloud utterance and increased the database to 437 records, or slightly more than 2 AUs per affective state manifested in the emote-aloud task.

Data Cleaning. There was a small number of observations for contempt (n = 8), curiosity (n = 3), and disgust (n = 5), so these affective states were not included in the current analysis. Anger and eureka had a non-trivial number of records (26 and 58 respectively), but they were excluded from this analysis because the associated records were not well distributed among the participants. We found that 88% of the records of anger were from only one participant. This data cleaning procedure resulted in reliable data only for boredom, confusion, and frustration. It is informative to note that boredom and confusion were the most frequent affective states in the Craig et al. (2004) study that had 34 participants. Therefore, we accepted the 5 participants’ data in the present study as reasonably representative of a larger population.
**Data Selection.** As would be expected, the database lacked an even distribution of the affective states among the 5 participants. Therefore, in order to achieve a uniform distribution of affective states, observations for each affective state were selected randomly from the AU database so that the data contained the same number of records per affective state for each participant. We selected 100 observations for each of the three affective states (boredom, confusion, frustration), which yielded 300 observations.

**Results & Discussion**

**Action Unit reliability**

Cohen’s Kappa tests of reliability were conducted on the coding of two trained judges for each clip with respect to the presence or absence of particular action units. Table 1 shows Kappa scores, segregated by each of the affective states, along with the proportion of observations that had each particular affective state. A mean Kappa of 0.76 was obtained by averaging the Kappa score over the participants. The mean Kappa was .84 for the affective states used for the data mining analyses (i.e., boredom, confusion, and frustration).

Insert Table 1 about here

Our raters’ high degree of reliability demonstrates the portability of Ekman’s Facial Action Coding System to affective states associated with learning (i.e., boredom, confusion, and frustration) as opposed to the basic emotions. The high reliability also justifies our search for meaningful patterns of AUs within our data. These data mining analyses focused on the three most frequent affective states in order to detect reliable patterns.

**Association rule mining**

A standard data mining procedure called the *a priori* algorithm (Agarwal & Srikant, 1994) was used to identify frequent sets of action units and to extract association rules that
could conditionally detect the presence of AUs on the face. Association rules are probabilistic in nature and take the form \( \text{Antecedent} \rightarrow \text{Consequent} \ [\text{support}, \text{confidence}] \). The antecedent is an AU or a set of AUs whose occurrence predicts the occurrence of the consequent (also an AU or a set of AUs). The support of a rule measures its usefulness and is the probability that the antecedent and the consequent occur simultaneously. The confidence measures its certainty and is the conditional probability that a data instance containing the antecedent will contain the consequent. For example, we observed an association rule with confusion of AU4 \( \rightarrow \) AU7 (See Table 2) for two action units AU4 (antecedent) and AU7 (consequent). This can be interpreted as “the presence of action unit 4 triggers action unit 7”. Its support is expressed as \( P[\text{AU4, AU7}] \), which is the proportion of records containing both action units 4 and 7. If \( P[\text{AU4}] \) is the proportion of records containing action unit 4, then the confidence of the rule is \( P[\text{AU4, AU7}] / P[\text{AU4}] \).

The association rule mining techniques yielded several relations between our three affective states and several unique action units. Table 2 highlights patterns composed of individual AUs that occur frequently, sets of AUs occurring together, and association rules of frequent AUs. The coverage for each pattern is the frequency of its presence in the 100 randomly selected data sets. Therefore, coverage of 100% for a pattern indicates that it was observed in all 100 data sets for an affective state. The average support is an average of the support values of all data sets (within an affective state) in which the pattern appeared.

Insert Table 2 about here

Our analyses were able to determine significant relationships with AUs for frustration, confusion, and boredom. It appears that AUs 1, 2, and 14 were primarily associated with frustration, but a strong association was found for a link between AUs 1 and
2 occurring together. Additionally, these AUs mutually trigger each other. That is, a raised inner brow tends to trigger a raised outer brow, and vice versa. Confusion displayed associations with AUs 4, 7, and 12. Action units 4 and 7 occur simultaneously and the presence of AU7 (tightened lids) tends to trigger AU4 (lowered brow). While boredom displayed a significant association with action unit 43 (eye closure), no association rules between action units were observed, but some weaker non-significant trends between eye movement, such as blinks and eye closure, and AUs related to mouth movement.

The AUs in Table 2 were associated with the designated affective states, but it is important to acknowledge that facial expressions are dynamic and change over time. Consequently, the AUs associated with a facial expression do not necessarily occur simultaneously. The association rule mining techniques also allow us to find auction units that temporarily co-occur within a particular affective state, but not all or most affective states. Notable exceptions are AUs 12 and 14 that occur during expressions of both confusion and frustration. Nevertheless, the strength of the discovered rules is derived from having a confidence of 100%. And finally, correlation analysis on the rules presented below ensures that the antecedent and consequent of each rule are positively correlated.

The AUs coded for each affective state received good kappa scores. Most of the AUs found to be important (i.e. 1, 2, 4, and 7) received a near perfect kappa score. This signifies that we can be fairly confident in the discriminability of the AUs in the detection of confusion and frustration. AU 12, AU 14, and AU 43, which received less support from the associations with the affective states, received lower, but acceptable, kappa scores.

Conclusions
The emote-aloud methodology in this study proved to be useful for studying affective states. This procedure allowed us to identify points during learning where affective events occurred. We were able to detect facial action units and patterns of action units that occur during the affective states of confusion, frustration, and boredom. Knowledge of the action units for these affective states has technological utility in addition to scientific value because it will allow us to reliably detect these affective states during learning. If these action units are automatically detected by sensing technologies, we have a solution to automatic detection of these affective states with facial feature recognition software (Cohn & Kanade, in press; Kapoor et al., 2001; Pantic & Rothkrantz, 2000).

This study is part of a larger research program that investigates the role of affect during the learning experience (D’Mello, Picard, & Graesser, in press). Our research program has three main objectives. The first is to identify the affective states that are most important during learning. Identification of these states requires an adequate understanding of theories of emotion and theories of learning. Our second objective is to find methods to reliably identify these affective states during learning. We are currently exploring non-intrusive sensing of affective states as learners interact with the AutoTutor program, such as the dialogue history logs, facial expressions, posture, and speech intonation. Finally, we will program AutoTutor to respond appropriately to affective states exhibited by learners. Learning is expected to improve if AutoTutor can adaptively respond to confusion, frustration, and boredom in the learner in addition to the learner’s cognitive states. For example, if the learner is frustrated, then AutoTutor should give a leading hint to put the student back on track. If the learner is bored, some engaging hooks are needed. If the learner is confused, there is an interesting pedagogical question of how long the student should
remain confused (and thinking deeply) before AutoTutor intervenes to put the student back on track.

The frequency with which affective states were reported could be one potential pitfall with this emote-aloud methodology. We found that 2 of 7 participants did not express affective states verbally. Four affective states were removed from the analysis due to floor effects: anger, disgust, contempt, and curiosity. Two alternative conclusions might be drawn from this. First, the reporting was low because the affective states do not occur during learning. This alternative would be compatible the observational study of Craig et al. (2004) on a larger sample of 34 participants. Second, participants might tend to under-report these states. Three of the four removed states were extreme negative affective states, so participants might be reluctant to report them. Nevertheless, the removal of anger, disgust, and contempt due to infrequent and inconsistent reporting reflects the observations made by a number of researchers that Ekman’s basic emotions are not particularly relevant to learning (Craig et al., 2004; Kapoor, et al., 2001).

Another potential drawback with our emote-aloud methodology addresses the consistency of participants’ reporting. The affective states were possibly reported haphazardly by some participants during the learning session and were not polled in set intervals. These unpredictable time samples prevent us from searching for stable sequential patterns of affective states that might be important during learning. We plan on pursuing other methodologies in the future while exploring the relations between affective states and complex learning.
References


 Cognition & Instruction, 19, 143-175.

 York: Cambridge University Press.


Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in student’s self- 
 regulated learning and achievement: A program of qualitative and quantitative 
 research. Educational Psychologist, 37, 91-105.


 confusion, concentration, and worry in an analysis of naturally occurring facial 

 emotional experience: The development and use of the Narcoder. In J.A. Coan, J.J. 
 University Press.

 and Bacon.

 Discourse Processes, 21, 225-286.
Acknowledgements

This research conducted by the authors was supported by the National Science Foundation (REC 0106965, ITR 0325428, REC 0633918) the Tutoring Research Group (visit http://www.autotutor.org). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF. We would also like to thank Laurentiu Cristofor for use of the ARMiner client server data mining application used for the association rule mining.
Table 1
Cohen’s Kappa between Two Coders and Proportion of the Data for each Emotion.

<table>
<thead>
<tr>
<th>Affective states</th>
<th>Cohen’s Kappa</th>
<th>Proportion of Emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>0.80</td>
<td>0.08</td>
</tr>
<tr>
<td>Boredom</td>
<td>0.84</td>
<td>0.20</td>
</tr>
<tr>
<td>Confusion</td>
<td>0.87</td>
<td>0.25</td>
</tr>
<tr>
<td>Contempt</td>
<td>1.00</td>
<td>0.04</td>
</tr>
<tr>
<td>Curiosity</td>
<td>0.50</td>
<td>0.01</td>
</tr>
<tr>
<td>Disgust</td>
<td>0.53</td>
<td>0.02</td>
</tr>
<tr>
<td>Eureka</td>
<td>0.86</td>
<td>0.14</td>
</tr>
<tr>
<td>Frustration</td>
<td>0.82</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Table 2

*Frequent Action Units along with Kappa Scores for Action Units (between the Two Raters Averaged across Participants) for 3 Frequent Affective States During Learning and Association Rules for each Affective state.*

<table>
<thead>
<tr>
<th>Affect</th>
<th>Pattern</th>
<th>Action Unit Description</th>
<th>Kappa Scores</th>
<th>Average Support</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frustration</td>
<td>2</td>
<td>Inner brow raise</td>
<td>0.93</td>
<td>20.2</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Outer brow raise</td>
<td>0.94</td>
<td>18.8</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>1,2</td>
<td>Inner and outer brow raised together</td>
<td>18.8</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1→2</td>
<td>Presence of an inner brow raise</td>
<td>18.8</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2→1</td>
<td>will trigger an outer brow raise and vice versa</td>
<td>19.2</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14*</td>
<td>Dimpler</td>
<td>0.82</td>
<td>11.6</td>
<td>66</td>
</tr>
<tr>
<td>Confusion</td>
<td>4</td>
<td>Brow lowerer</td>
<td>1.00</td>
<td>23.9</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Lid tightener</td>
<td>0.99</td>
<td>20.2</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>4, 7</td>
<td>Brow lowered with tightened lids</td>
<td>18.6</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7→4</td>
<td>Tightened lids will lead to a lowered brow</td>
<td>18.6</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12*</td>
<td>Lip corner puller</td>
<td>0.70</td>
<td>18.5</td>
<td>95</td>
</tr>
<tr>
<td>Boredom</td>
<td>43*</td>
<td>Eye closure</td>
<td>0.77</td>
<td>23.9</td>
<td>40</td>
</tr>
</tbody>
</table>

*Notes: → implies an association rule with 100% confidence

* Of secondary importance