The Affective Impact of Tutor Questions: Predicting Frustration and Engagement

Alexandria K. Vail  
Department of Computer Science  
North Carolina State University  
Raleigh, North Carolina  
akovail@ncsu.edu

Joseph B. Wiggins  
Department of Computer Science  
North Carolina State University  
Raleigh, North Carolina  
jbwigg3@ncsu.edu

Joseph F. Grafsgaard  
Department of Psychology  
North Carolina State University  
Raleigh, North Carolina  
jgrafsg@ncsu.edu

Kristy Elizabeth Boyer  
Department of Computer & Information Science & Engineering  
University of Florida  
Gainesville, Florida  
keboyer@ufl.edu

Eric N. Wiebe  
Department of STEM Education  
North Carolina State University  
Raleigh, North Carolina  
wiebe@ncsu.edu

James C. Lester  
Department of Computer Science  
North Carolina State University  
Raleigh, North Carolina  
lester@ncsu.edu

ABSTRACT

Tutorial dialogue is a highly effective way to support student learning. It is widely recognized that tutor dialogue moves can significantly influence learning outcomes, but the ways in which tutor moves, student affective response, and outcomes are related remains an open question. This paper presents an analysis of student affective response, as evidenced by multimodal data streams, immediately following tutor questions. The findings suggest that students’ affect immediately following tutor questions is highly predictive of end-of-session self-reported engagement and frustration. Notably, facial action units which have been associated with emotional states such as embarrassment, disgust, and happiness appear to play important roles in students’ expressions of frustration and engagement during learning. This line of investigation will aid in the development of a deeper understanding of the relationships between tutorial dialogue and student affect during learning.

Keywords

Tutorial dialogue, affect, frustration, engagement, facial expression

1. INTRODUCTION

Tutorial dialogue provides rich, natural language adaptation to students during learning. An understanding has emerged about the role of interactivity in tutorial dialogue [40, 6] and on dialogue strategies for most effectively supporting students in task-oriented tutorial dialogues [29, 10].

However, a pressing issue is developing an understanding of how specific tutor dialogue moves impact students’ affect, and in turn, what influence students’ affective responses may have on outcomes.

The need for modeling affect during learning is widely recognized. Research has shown that suites of affect detectors from sensors and log files can perform well but that there are trade-offs depending on the goals of the affect detection modules [22, 33]. Affect detectors have been investigated for a wide variety of affective states including confidence, excitement, frustration, and interest [41], and within tutorial dialogue, for uncertainty [11]. There have also been great strides in sensor-free affect detection which relies primarily on log files [2]. This approach has shown promise during cognitive tutoring [9] and for distinguishing frustration and confusion [27].

Out of all of the affective phenomena that have been examined during learning, two affective states are frustration and engagement. These states have been examined in fine-grained analyses as tutoring unfolds, and also as outcome measures regarding students’ perceptions of the success of the tutoring session. Engagement and frustration have been predicted at above-chance levels using facial expression-based affect detection even without the presence of interactive events during text or diagram comprehension [5]. Engagement and frustration have also been predicted with nonverbal behaviors, including facial expression, after student task events during problem solving [16]. In a compelling development, emerging evidence shows that fine-grained affective events can have long-lasting relationships with outcomes that may be far removed from those affective events [36].

This paper advances the understanding of student emotions in learning by examining students’ fine-grained affective responses to tutor questions during tutorial dialogue. It investigates the hypothesis that students’ affective responses immediately following tutor questions are related to self-reported frustration and engagement at the end of the session. The results indicate that several key facial expression
features immediately following two different types of tutor questions are highly predictive of end-of-session self-reported engagement and frustration. This line of investigation represents a step forward in understanding the affective impact of tutorial strategies.

2. RELATED WORK

Tutorial dialogue researchers have long studied what human tutors naturally do: how strategies differ between experts and novice tutors [12] whether Socratic or didactic approaches are most effective [35] and how tutors scaffold and fade support during problem solving [4], among others. The impact of particular tutorial dialogue moves has been the focus of significant attention, with findings indicating that positive and negative feedback have different impact based on students’ self-efficacy level [3], that bottom-out directives are not conducive to learning [29], and that adapting to student uncertainty improves the effectiveness of tutorial dialogue [10]. However, this paper examines a different aspect of these tutorial dialogue moves that is critical in learning: students’ affective response as expressed on the face and as embodied in gestures.

Multimodal features such as dialogue, facial expression, posture, and task actions have been used to predict affective states, such as boredom, confusion, excitement, and frustration, as those states occur during learning [23, 8, 7]. Moreover, multimodal features such as facial expression and gestures can significantly predict frustration and engagement reported at the end of tutoring sessions [17], and some differences have emerged in the extent to which upper and lower facial expression features are associated with these outcomes [15]. This previous work on utilizing multimodal features for predicting frustration and engagement during human-human tutoring has emphasized the important role that tutor dialogue moves play in affective outcomes. Other factors, such as student personality profile, can also contribute significantly to predicting these outcomes [39]. The present work examines moment-by-moment affect as evidenced by multimodal traces, and then analyzes the relationship between these multimodal behaviors and the outcomes of frustration and engagement as reported by students after the tutoring session.

3. STUDY DATA

The present analysis investigates the multimodal behavior of students during a computer-mediated tutorial session in introductory computer science, and specifically in Java programming [18, 30]. The tutorial interface, shown in Figure 1, is divided into four panes: the task description, the student’s Java source code, the compilation and execution output of the program, and the textual dialogue messages between the tutor and the student. The tutor’s interactions with the environment were constrained to progression between tasks and sending textual messages to the student.

Students (N = 67) were university students in the United States enrolled in an introductory engineering course, with an average age of 18.5 years (s = 1.5 years), whereas the human tutors (N = 5) were primarily graduate students with previous experience in tutoring or teaching introductory programming. The behavior of the student was collected using a set of multimodal sensors, as shown in Figure 2, including a Kinect depth sensor, an integrated webcam, and a skin conductance bracelet. The following subsections detail the modalities appearing significant in the present analysis.

Each student participated in six 40-minute sessions over the course of four weeks; however, the present analysis only examines data from the first lesson. Before and after each lesson, students completed a content-based pretest and identical posttest; the tutoring sessions were found to be significantly effective in facilitating learning gains (p ≪ 0.0001). In addition to the posttest, students also completed a post-survey, including the NASA-TLX workload survey [20] and the User Engagement Survey [32]. The present analysis investigates self-reported frustration, taken from the Frustration Level item of the NASA-TLX workload survey, and engagement, taken as an average of three sub-scales of the User Engagement Survey: Focused Attention (perception of time passing), Felt Involvement (perception of involvement with the session), and Endurability (perception of the activity as worthwhile).

3.1 Task Event and Dialogue Features

During the tutoring session, the interface described above logged tutor and student dialogue messages, student typing in the code window, and student progress through the task. No turn-taking measures were enforced in the dialogue: students and tutors could send messages to the other at any point. All exchanged messages were automatically tagged by a J48 decision tree classifier [37] with a dialogue act annotation scheme created for task-oriented tutorial dialogue that differentiates tutor questions, feedback, and hints, among other dialogue moves [38]. In that work, the Cohen’s kappa between two human annotators was 0.87 and the Cohen’s kappa between human and the J48 decision tree classifier was 0.786.

The analysis presented here focuses on two types of tutor dialogue moves: inference questions and evaluative questions. (Although other question types were investigated, student reactions to these were not found to have significant predictive power.) Inference questions require the formation of an action plan or reasoning about existing content knowledge. For example, ‘How do you think this problem can be solved?’, or ‘How can you fix this error?’ are considered to be inference questions. On the other hand, evaluative questions aim to evaluate the student’s belief in his or her own understanding of the material, e.g., ‘Does that make sense so far?’, or ‘Do you understand?’ (see Figure 4). Previous work has suggested that questions can stimulate cognitive disequilibrium in a student [34], which is often considered to be a critical step in knowledge acquisition [13]. On the other hand, evaluative questions that ask a novice to evaluate whether she understands material may not be particularly helpful pedagogically because novices often cannot identify what they do not understand, or may be hesitant to speak up even if they are aware that they are confused. Nonetheless, these questions occurred regularly in our corpus with experienced (though not expert) human tutors. We investigate whether students’ affective response to these types of tutor dialogue moves is significantly predictive of student engagement and frustration as reported at the end of the session.

3.2 Facial Expression Features

Student facial expressions were automatically extracted
using a state-of-the-art facial expression recognition tool-
box, FACET (commercial software preceded by a research
version known as the Computer Emotion Recognition Tool-
box, CERT) [26]. FACET tracks the frame-by-frame pres-
ence of several facial action units according to the Facial
Action Coding Scheme [25]. These action units include
movements such as AU6 CHEEK RAISER, AU12 LIP COR-
NER PULLER, AU24 LIP PRESSOR, and AU26 JAW DROP
(see Figures 5 and 6 for illustration). For each facial action
unit, the FACET software suggests an Evidence
measure, indicating the chance that the target expression is present.
This Evidence measure is on a scale where negative values
represent evidence of the absence of a facial expression and
positive values indicate evidence of the presence of one. The
more positive the measure, the more confident FACET is
that the feature is present.

3.3 Gesture Features
The Kinect depth camera also tracked hand-to-face ges-
tures made by the student during the tutoring session. An
algorithm developed to detect such gestures was developed
to recognize one or two hands touching the lower face. In
order to do this, the algorithm relies on surface propagation
from the center of the head, identifying round (i.e. a normal
head shape) or oblong shapes (i.e., shapes extending beyond
the normal head shape) based on distances from the center
of the head. This gesture detection algorithm was previously
found to be 92.6% accurate when compared against manual
labels [14].

4. ANALYSIS
The present analysis focuses on the affective response of
a student, as observed by multimodal traces of face and
gesture, after tutor inference questions and evaluative ques-
tions. We hypothesize that multimodal features after these
tutor questions can predict student engagement and frustra-
tion. In particular, we examine three seconds after each
tutor dialogue move (a manually-determined interval). The
multimodal response of the student was characterized using
the following categories of features, all of which were pro-
vided to the predictive models. However, note that only the
first two of these categories of features (shown in bold below)
appear significantly predictive within the models.

1. Average evidence measure for each of the facial
expression action units during the interval (19 features)

2. Percentage of the interval in which a one-hand-
to-face or two-hands-to-face gesture was observed
(2 features)

3. Number of skin conductance responses identified dur-
ing the interval as measured by a skin conductance
response bracelet (1 feature)

4. Average student distance from the workstation during
the interval (1 feature)

5. Average difference between the highest and lowest points
of the student’s body from the workstation during the
interval, indicating leaning (1 feature)

We calculated the average value of each multimodal fea-
ture listed in the categories above across each tutoring ses-
tion. For each feature, we computed its conditional proba-
bility of occurring after the tutor moves of inference question
or evaluative question. We also provided the model with the
overall occurrence of that feature across the entire tutoring
Figure 2: Multimodal instrumented tutoring session, including a Kinect depth camera to detect posture and gesture, a webcam to detect facial expression changes, and a skin conductance bracelet to detect electrodermal activity.

Figure 3: Dialogue excerpt illustrating a tutor inference question in context.

\textbf{Student compiles the program, encounters an error.}

\begin{tabular}{|p{3cm}|p{7cm}|}
\hline
\textbf{Student} & Oh. \\
\hline
\textbf{Tutor} & So how can we fix this? \\
\hline
\textbf{Student} & Hmm. \\
\hline
\textbf{Student} & Switch the prompt line with the response line? \\
\hline
\textbf{Tutor} & Okay, try it. \\
\hline
\end{tabular}

Figure 4: Dialogue excerpt illustrating a tutor evaluative question in context.

\begin{tabular}{|p{3cm}|p{7cm}|}
\hline
\textbf{Student} & Do I need to set the player input before line 13? \\
\hline
\textbf{Tutor} & The \textbf{while} tests that [variable]. You need to be sure it enters the loop at least once. \\
\hline
\textbf{Tutor} & Good. \\
\hline
\textbf{Tutor} & Does that make sense? \\
\hline
\textbf{Student} & Yeah. \\
\hline
\textbf{Student} & But what happens if I don’t enter 1 or 2? \\
\hline
\end{tabular}

A session in order to control for the influence of the feature overall (rather than only after the tutor moves of interest). Specifically, the features conditional on tutor moves were averages of the form $\text{Avg}(\text{Feature}|\text{TutorQ})$ for each student that completed the session. The session-wide average of each feature, $\text{Avg}(\text{Feature})$ were also provided to the model for each multimodal feature in all of the categories above.

Standardization was performed on each feature by subtracting the mean and dividing by the standard deviation, so that the regression coefficients would be more interpretable. The standardized features were provided to a stepwise regression modeling procedure optimizing for the leave-one-student-out cross-validated $R^2$ value (the coefficient of determination), while at the same time requiring a strict $p < 0.05$ cut-off value after Bonferroni correction on significance values.

5. RESULTS AND DISCUSSION

For both types of tutor question, evaluative and inference, a predictive model was built to predict student frustration and student engagement, resulting in a potential four models. Three of the four models uncovered significant predictive relationships. The following subsections detail models predicting frustration after tutor inference and evaluative questions, and a model predicting engagement after tutor evaluative questions.

5.1 Frustration

The results suggest that student facial expressions are significantly predictive of self-reported end-of-session frustration. The predictive model for student frustration based on tutor evaluative questions includes two features, both of which are facial action units occurring in the three-second interval following the tutor evaluative question (Table 1). Two facial action unit features after tutor evaluative questions a part of a larger exploratory analysis. As a result, the $p$-values reported have been modified by a Bonferroni correction.
Table 1: Predictive model for standardized end-of-session frustration after tutor evaluative questions (TutorQE).  

<table>
<thead>
<tr>
<th>Frustration =</th>
<th>$R^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-0.7039 \times \text{AU12 after TutorQE}$</td>
<td>0.0764</td>
<td>0.014</td>
</tr>
<tr>
<td>$-0.6279 \times \text{AU28 after TutorQE}$</td>
<td>0.2471</td>
<td>0.030</td>
</tr>
<tr>
<td>$-0.1635$ (Intercept)</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Leave-One-Out Cross-Validated $R^2 = 0.3235$

Table 2: Predictive model for standardized end-of-session frustration after tutor inference questions.

<table>
<thead>
<tr>
<th>Frustration =</th>
<th>$R^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$+0.5660 \times \text{AU6 after TutorIQ}$</td>
<td>0.2893</td>
<td>0.022</td>
</tr>
<tr>
<td>$+0.3635 \times \text{AU20}$</td>
<td>0.0499</td>
<td>0.019</td>
</tr>
<tr>
<td>$-0.0174$ (Intercept)</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Leave-One-Out Cross-Validated $R^2 = 0.3392$

Predictive questions. For inference questions, none of the features provided to the model were predictive of engagement. However, for affective response to tutor evaluative questions, there were seven predictive features, three of which are specific to the interval following the event, and four of which are session-wide (Table 3).

The model suggests that facial expression features account for most of the variance in predicting student engagement; however, one session-wide gesture feature was also
Table 3: Predictive model for standardized engagement after tutor evaluative questions.\(^1\)

<table>
<thead>
<tr>
<th></th>
<th>(R^2)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+0.4422 * ONEHTF</td>
<td>0.1815</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>−0.5989 * AU10 after TUTOREQ</td>
<td>0.1831</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>+0.5770 * AU12</td>
<td>0.2280</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>+0.5097 * AU26 after TUTOREQ</td>
<td>0.0514</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>−0.2941 * AU2</td>
<td>0.1923</td>
<td>0.003</td>
</tr>
<tr>
<td>+0.2467 * AU5</td>
<td>0.0295</td>
<td>0.002</td>
</tr>
<tr>
<td>+0.1792 * AU24 after TUTOREQ</td>
<td>0.0566</td>
<td>0.018</td>
</tr>
<tr>
<td>+0.4100 (Intercept)</td>
<td></td>
<td>1.000</td>
</tr>
</tbody>
</table>

Leave-One-Out Cross-Validated \(R^2 = 0.9224\)

selected. The more frequently a student was displaying a OneHandToFace gesture, which may indicate thoughtful contemplation, the more engaging the student reported the experience at the end of the session.

Three more session-wide facial expression features were selected as significantly predictive of student engagement. The more intense the expression of AU12 Lip Corner Puller (Figure 5b) or AU5 Upper Lid Raiser (Figure 6b), the more engaged the student. For AU12 which is often associated with smiling, a positive emotion is likely related to higher engagement. In this task, AU5 is likely associated with the student looking at the screen, possibly indicating paying attention and focusing on the task (as opposed to the opposite facial movement of blinking or shutting one’s eyes). In contrast, AU2 Outer Brow Raiser (Figure 6a) was predictive of lower engagement. This action unit is a component of the “fear brow” (AU1+2+4) which has been evidenced as a display of anxiety [19].

Narrowing down to the context of three seconds after tutor evaluative questions, three facial expression features were significantly correlated with student engagement. The more that a student expresses AU26 Jaw Drop (Figure 6c), or the more that the student expresses AU24 Lip Pressor (Figure 6d), the more engaged the student reported being at the end of the session. Jaw drop is a dynamic action unit that may occur when the mouth is closed or already partly open. In either case, this action unit may be associated with focus on the task, although it could also plausibly be associated with a yawn (which we would not expect to coincide with higher engagement). With respect to AU24, which is a prototypical component of anger, an important interplay of learning and affect expression emerges. Some facial movements that are part of prototypical displays of negative basic emotions, such as anger, appear to be indicative of mental effort during learning, rather than negative affect [31]. From this perspective, it makes sense that this AU24 would be related to engagement. On the other hand, the more that a student expressed AU10 Upper Lip Raiser (Figure 6c) during this interval, the less engagement reported by the student at the end of the session. This action unit, which is a component of prototypical disgust, is likely to run contrary to engagement.

6. CONCLUSION

Tutor dialogue moves in one-on-one human tutoring significantly influence student outcomes, both cognitive and affective. This paper has examined students’ affective response to two types of tutor questions: inference questions which require some reasoning to construct an answer, and evaluative questions, which ask students to reflect on the extent to which they understand the material. The results show that immediately after these tutor questions, students’ affective displays—particularly with respect to facial expression—are highly predictive of the outcomes of frustration and engagement. By detecting these affective displays which have been associated in prior studies with emotions such as embarrassment, disgust, or happiness, we can begin to understand the moment-by-moment affective processes that influence learning through tutorial dialogue, and relate those fine-grained events to overall outcomes.

While these facial movements have been associated with prototypical emotion displays in the literature, it is important to further contextualize the moments in which these expressions appear during tutoring. For instance, action units typically associated with anger are likely indicators of mental effort during learning. Similarly, an action unit associated with disgust (e.g., AU10) may be related to students’ appraisal of the tutor’s question in the moment. Further research seeks to ground these interpretations more extensively across salient moments of tutoring.

There are several additional directions for future work. Detecting important moments during tutoring is an open area of investigation, with evidence suggesting that moment-by-moment affect may be related to distal outcomes [36, 1]. In future work, it will be important to expand our understanding of the identified non-verbal predictors for frustration and engagement more deeply. We must consider a wider variety of contexts, and explore different widths of time after tutorial events to examine affective responses with longer (or shorter) times to manifest. It is hoped that this line of investigation will lead to richer affect models for tutorial dialogue.

Acknowledgments

The authors wish to thank the members of the LearnDialogue and Intellimedia groups at North Carolina State University for their helpful input. This work is supported in part by the Department of Computer Science at North Carolina State University and the National Science Foundation through Grants IIS-1409639, CNS-1453520, and a Graduate Research Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the participants, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

7. REFERENCES


[29] C. M. Mitchell, E. Y. Ha, K. E. Boyer, and J. C. Lester. Learner characteristics and dialogue: recognising effective and student-adaptive tutorial


Figure 6: Sample frames from the student webcam illustrating the facial action unit features appearing in the predictive model for student engagement, as identified by FACET. Note that AU12 LIP CORNER PULLER (Figure 5b) also appears in these models.