Enhancing Student Models in Game-based Learning with Facial Expression Recognition

Robert Sawyer  
Department of Computer Science  
North Carolina State University  
rssawyer@ncsu.edu

Andy Smith  
Department of Computer Science  
North Carolina State University  
pmsmith4@ncsu.edu

Jonathan Rowe  
Department of Computer Science  
North Carolina State University  
jprowe@ncsu.edu

Roger Azevedo  
Department of Psychology  
North Carolina State University  
razved@ncsu.edu

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

ABSTRACT
Recent years have seen a growing recognition of the role that affect plays in learning. Because game-based learning environments elicit a wide range of student affective states, affect-enhanced student modeling for game-based learning holds considerable promise. This paper introduces an affect-enhanced student modeling framework that leverages facial expression tracking for game-based learning. The affect-enhanced student modeling framework was used to generate predictive models of student learning and student engagement for students who interacted with CRYSTAL ISLAND, a game-based learning environment for microbiology education. Findings from the study reveal that the affect-enhanced student models significantly outperform baseline predictive student models that utilize the same gameplay traces but do not use facial expression tracking. The study also found that models based on individual facial action coding units are more effective than composite emotion models. The findings suggest that introducing facial expression tracking can improve the accuracy of student models, both for predicting student learning gains and also for predicting student engagement.

KEYWORDS
Student Modeling, Affect, Game-based Learning

1 INTRODUCTION
Affect plays a central role in learning [3, 8, 9, 48]. Positive emotions, such as curiosity and joy, can lead students to engage more deeply in learning, while emotions such as boredom and frustration can lead students to resort to unproductive behaviors including hint abuse and gaming the system, or even to completely disengage [37, 48]. Students exhibit a range of emotions when interacting with adaptive learning environments [13]. Building on recent advances in our understanding of affect, which stem from both manually coded protocols [15, 35] and automated techniques [7, 10, 36], affect-enhanced student models that track and dynamically respond to affect on a moment-by-moment basis offer significant potential for enhancing adaptive learning environments.

While student modeling has long been a goal of the user modeling community [19, 33], student modeling work to date has focused primarily on cognitive aspects of learning [1]. Recent student modeling work has begun to consider how to augment student models with affect [12], as well as to tailor support based on students’ affective states, such as confusion, frustration, and boredom [14, 17]. Models of affect have also been used to predict students’ reactions to virtual agents [25, 29] and better understand student motivation and off-task behavior [40].

Game-based learning has been studied in a broad range of subject matters and student populations [11, 20, 42, 49]. Because game-based learning environments are designed to provide students with engaging learning experiences, they offer a rich “laboratory” for investigating affect-enhanced student models. This paper reports on an investigation of affect-enhanced student modeling for game-based learning. Specifically, using a game-based learning environment for microbiology education, CRYSTAL ISLAND [38], we compare two approaches to student modeling: an affect-enhanced student model that uses facial expression tracking in addition to gameplay data, and a baseline student model that uses only gameplay data. We compare the accuracy of the models for predicting both student learning (measured with normalized learning gain) and student presence (a facet of engagement measured with participant perception of transportation into a virtual environment through a standard self-report instrument [4]). In addition, we compare two competing approaches to affect recognition—composite emotion models and models based on individual facial action coding units—to explore how different types of facial expression tracking can support affect-enhanced student modeling.
2 RELATED WORK

Learning and affect are inextricably linked. It has been found, for example, that positively valenced emotions tend to support student learning [37]. Confusion has been shown to be beneficial to learning in certain contexts [16], yet students experiencing anxiety or anger tend to be less efficient learners [37]. Student affect correlates with problem-solving performance [39] and is predictive of user interactions with virtual agents [25]. Recognizing the potential for affect to inform student modeling in learning environments, recent work with automated affect detectors has found that integrating detectors into Bayesian knowledge tracing models can improve predictive accuracy of student performance [12].

Although there are many open questions about which emotions or sets of emotions offer the greatest potential for adaptive learning environments, as well as which sets of multimodal data streams can best support adaptivity in learning environments, facial expression tracking has emerged as a promising means for inferring learning-centric emotions [10, 28]. Arroyo et al. [2] used facial expressions in combination with other sensors to trigger interventions, which appeared to increase on-task behavior and reduce hint-abuse. Bosch et al. [7] created models that accurately predict student affect with facial action units in a classroom setting that generalized across multiple days and multiple classrooms. Grafsgaard et al. [24] found an additive relationship between facial expressions, dialogue acts, and task performance in a tutoring system, and more recent work in this line of investigation found that male and female students exhibit significant differences in facial expressions [43]. Collectively, this work suggests that facial expressions provide a window into affect and learning that can be utilized by student models to support adaptive learning environments.

Upon detecting student affective states, adaptive learning environments can deliver a broad range of possible supports for student learning and emotion regulation. For example, D’Mello et al. [14, 17] devised a modified version of AutoTutor that delivered affect-sensitive dialogue moves, facial displays of emotion, and synthesized speech patterns in response to student affect. Results indicated that the affect-sensitive interventions had a positive impact on student learning among low prior knowledge students, but a neutral or even harmful impact on student learning in other conditions. DeFalco et al. [18] found that self-efficacy-oriented feedback messages, which were delivered in response to detected learner frustration, helped to enhance student learning in a simulation-based training environment for military combat medics. However, the feedback messages were not found to mitigate occurrences of frustration itself during training.

Although student modeling has been investigated for decades in the intelligent tutoring systems community [44], only recently has student modeling research turned to game-based learning. Baker et al. [6] analyzed game-trace logs in the Virtual Performance Assessments system to create assessment models of students’ ability to design controlled experiments and apply inquiry skills. Using evidence-centered design, Bayesian network-based stealth assessments of student knowledge have been explored in both the Physics Playground and a modified version of the popular game, Plants vs. Zombies [26, 41], and other work in this vein has explored deep learning-based models of stealth assessment for game-based learning [31]. Student goal recognition has also been investigated in game-based learning, with approaches spanning Bayesian networks [34], Markov logic networks [5], and long short-term memory networks [32]. The affect-enhanced student modeling approach presented here extends earlier work in student modeling for game-based learning to exploit facial expression data to improve student models’ predictive accuracy.

Figure 1. Screenshots from CRYSTAL ISLAND illustrating a virtual textbook, non-player characters, the virtual scanner, and game environment.
3 CRYSTAL ISLAND

We explore affect-enhanced student modeling in a study conducted with an intelligent game-based learning environment for science problem solving in microbiology, CRYSTAL ISLAND [38]. Intelligent game-based learning environments integrate the personalized learning supports of intelligent tutoring systems [27, 45] with the engaging interactions afforded by game technologies. In CRYSTAL ISLAND, students take on the role of a medical field agent tasked with investigating an epidemic on a remote island research station. The student must discover the source and identity of the disease as well as recommend a treatment plan to aid the infected members of the island’s research team.

In CRYSTAL ISLAND, students engage in a range of problem-solving actions to gather information, test hypotheses, and diagnose a solution. Examples of student activities in the game include exploring the island environment, conversing with non-player characters, running tests on potentially contaminated items in a virtual laboratory, and completing an in-game diagnosis worksheet to report findings and solve the mystery. The open-world game environment features several buildings, including an infirmary, dining hall, laboratory, and various residences where the student can gather information by talking to characters. Relevant microbiology concepts about viruses, bacteria, immunization, and how diseases spread are found in virtual books and articles scattered throughout the buildings of the island. Students use knowledge gained from both the informational texts and dialogue with non-player characters to diagnose the illness and solve the mystery. Students successfully complete the game by submitting a worksheet to the camp nurse with the correct diagnosis, identified transmission object, and treatment solution.

4 METHODS AND DATA

To investigate the potential contributions of affect to student modeling, we compare two approaches. First, we created an affect-enhanced student model that uses facial expression tracking in combination with gameplay. As students played CRYSTAL ISLAND, video was captured of their facial expressions. We created predictive student models that used both the facial expression data stream and the gameplay data stream to predict students’ learning and students’ sense of presence, a core component of engagement. We then created a baseline student model that also predicts students’ learning and presence but only uses the gameplay data stream. We compare the performance of these two models and explore the impact of two approaches to facial recognition tracking for affect-enhanced student modeling: (1) composite emotion models and (2) models based on individual facial action coding units.

4.1 Participants and Experimental Setup

The study involved 37 college age students that interacted with the CRYSTAL ISLAND game-based learning environment in a lab setting. Four students were removed due to partial or missing data, resulting in 33 students ($M = 19.9$ years old, $SD = 1.30$), of which 17 (51.5%) were female. All remaining students played the game to completion, and time spent playing the game ranged from 26.4 to 105.1 minutes ($M = 63.7$, $SD = 18.4$). Prior to interacting with the game, students completed a 20-question multiple-choice test assessing their conceptual and application-based understanding of microbiology. At conclusion of gameplay, students completed a Presence Questionnaire [46] and again completed the same microbiology assessment.

4.2 Survey Measures

In order to assess student learning, we examined students’ performance on the microbiology content test administered before and after students’ interactions with CRYSTAL ISLAND. Using the difference between the pre-test score (Pre-Test) and post-test score (Post-Test), Normalized Learning Gain was calculated for each student participating in the study. Normalized Learning Gain is the difference between Post-Test and Pre-Test, standardized by the total amount of improvement or decline possible from the Pre-Test. Students achieved positive Normalized Learning Gain on average ($M = 0.226$, $SD = 0.315$), with 25 of the 33 (75.8%) students achieving positive Normalized Learning Gain (Post-Test greater than Pre-Test).

$$\text{Normalized Learning Gain} = \begin{cases} \frac{\text{Post} - \text{Pre}}{1 - \text{Pre}} & \text{Post} > \text{Pre} \\ \frac{\text{Post} - \text{Pre}}{\text{Pre}} & \text{Post} \leq \text{Pre} \end{cases}$$

We use presence as a proxy for engagement in the game environment. In order to assess student presence, the Presence Questionnaire was used to measure players’ self-reported perceptions of presence [47], which refers to a participant’s perception of transportation into a virtual environment. The questionnaire contains a series of 30 questions that students answer on a 7-point scale to characterize their in-game experience with a virtual environment. The questionnaire consists of several subscales for measuring presence, such as involvement, sensory fidelity, adaptation/immersion, and interface quality. To measure presence, we summed the numeric responses to all 30 questions to obtain a presence score for each student, which ranged from 116 to 209 ($M = 155$, $SD = 23.4$).

4.3 Facial Expression Recognition

Facial expression features were extracted automatically through a video-based facial expression tracking system, iMotions. The iMotions software was previously commercially available as FACET and the research-focused toolbox CERT [30].
evidence scores for 20 AUs and 9 composite affective measures. The 9 composite affective measures include *Anger*, *Surprise*, *Frustration*, *Joy*, *Confusion*, *Fear*, *Disgust*, *Sadness*, and *Contempt*. Table 1 displays the relationships between composite affective measures and AUs as defined by iMotions.

**Table 1. Breakdown of Composites by contributing AU.**

<table>
<thead>
<tr>
<th>Composite Measure</th>
<th>Contributing AUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jay</td>
<td>6 + 12</td>
</tr>
<tr>
<td>Sadness</td>
<td>1 + 4 + 15</td>
</tr>
<tr>
<td>Surprise</td>
<td>1 + 2 + 5 + 26</td>
</tr>
<tr>
<td>Fear</td>
<td>2 + 4 + 5 + 7 + 20 + 26</td>
</tr>
<tr>
<td>Anger</td>
<td>4 + 5 + 7 + 23</td>
</tr>
<tr>
<td>Disgust</td>
<td>9 + 15 + 16</td>
</tr>
<tr>
<td>Contempt</td>
<td>12 + 14</td>
</tr>
</tbody>
</table>

Predictive models require higher-level features than are provided by the evidence scores. We performed feature engineering to transform the evidence scores to a representation that could be effectively used by the student models. We used a method with relative thresholding by amplitude and summed over durations to minimize the effect of micro expressions. First, the evidence scores were standardized for each student by subtracting the mean evidence and dividing by the standard deviation evidence over their entire episode. While iMotions calibrates the evidence scores over the first few observations, it does not account for the potential variability of evidence for students. Thus, dividing by the standard deviation of a student’s evidence score over the episode accounts for potential variability in expressiveness of individuals. Events were added to the affect log for the duration that the standardized evidence scores rose above a threshold of one. These events represent moments when evidence scores rose one standard deviation above the mean evidence score for a particular student. The duration of these events in each affective measure (i.e., the 20 AUs and 9 Composites) is summed over the gameplay episode of a student. The final feature used in student modeling is this sum (i.e., the duration of the evidence score was above one standard deviation above the mean evidence score) divided by the total time the student spent playing the game. These affective features thus represent the proportion of gameplay duration that the respective affective evidence score was one standard deviation above the mean evidence score for a particular student.

### 4.3 Gameplay Behavior Logging

Gameplay interactions with **CRYSTAL ISLAND** were recorded in a granular timestamped game log for each student. From these game interaction logs, several measures were calculated to summarize student behavior in the game-based learning environment. The actions recorded include how students gather information (e.g., reading books and articles, speaking with non-player characters), select and organize information (e.g., editing the diagnosis worksheet), and test their hypotheses (e.g., scanning objects, submitting the worksheet as the final solution). The

---

**Figure 2. Examples of AUs used.**

measures used as features in predictive modeling include both the number of actions performed and the duration of actions with varying lengths. The set of actions reported includes conversations with non-player characters (*ConversationCount* and *ConversationDuration*), items scanned in the virtual laboratory (*ScannerCount*), books and articles read (*BooksReadCount* and *ReadingDuration*), edits to the diagnosis worksheet (*WorksheetCount* and *WorksheetDuration*), and the number of times the worksheet was submitted for final evaluation (*SubmitCount*). These measures were also standardized by the total time the student spent playing the game to represent rate-per-minute for counts and proportion of time spent performing each of the actions over a given duration (Table 2). The units for counts are shown in counts-per-minute, while the units for durations are in seconds-per-minute, representing a proportion out of 60.

**Table 2. Mean and standard deviations of gameplay behaviors.**

<table>
<thead>
<tr>
<th>Gameplay Behaviors</th>
<th>Mean (per minute)</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ConversationCount</em></td>
<td>0.797</td>
<td>0.216</td>
</tr>
<tr>
<td><em>BooksReadCount</em></td>
<td>0.352</td>
<td>0.119</td>
</tr>
<tr>
<td><em>WorksheetCount</em></td>
<td>0.352</td>
<td>0.159</td>
</tr>
<tr>
<td><em>SubmitCount</em></td>
<td>0.029</td>
<td>0.017</td>
</tr>
<tr>
<td><em>ScannerCount</em></td>
<td>0.389</td>
<td>0.223</td>
</tr>
<tr>
<td><em>ConversationDuration</em></td>
<td>7.900</td>
<td>2.070</td>
</tr>
<tr>
<td><em>ReadingDuration</em></td>
<td>24.100</td>
<td>5.280</td>
</tr>
<tr>
<td><em>WorksheetDuration</em></td>
<td>5.150</td>
<td>1.970</td>
</tr>
</tbody>
</table>

### 5 RESULTS

We used three different subsets of the full feature set in order to compare predictive student models of Normalized Learning Gain and Presence. The first feature subset (Gameplay) consisted of only the 8 gameplay features (actions and durations of select actions) to represent a baseline model without affect information. The second feature subset consisted of the gameplay actions and
the 9 different composite affective features (Gameplay+Composites). The third feature subset consisted of the gameplay actions with the 20 different AU affective features (Gameplay+AUs). The composite features and AUs are treated separately due to multicollinearity issues stemming from the fact that composites are combinations of the AUs. The AUs provide a more granular measure while the composites provide a higher level, more interpretable measure of affect. We compare the performance of each feature subset predicting Normalized Learning Gain and Presence for a total of 6 models (3 feature subsets x 2 response variables), with 2 being baseline models using only gameplay features, and the other 4 being affect-enhanced models of student learning and presence.

Variables for a linear model were chosen using a forward stepwise linear regression where features were selected from their respective subset to optimize leave-one-out cross-validation (LOOCV) $R^2$. The reported models are the result of using the selected features in an ordinary least squares linear regression model.

### 5.1 Baseline Models

Baseline predictive models for Normalized Learning Gain and Presence were created using the stepwise feature selection process using the subset of 8 gameplay features. In addition to standard linear regression measures, (coefficients, standard error of coefficients, standardized coefficients and t-statistic for significance of feature), the leave-one-out cross validation $R^2$ is reported as the measure of the accuracy of the model.

The highest performing Normalized Learning Gain model using gameplay features for leave-one-out cross validation $R^2$ is shown in Table 1. This linear model uses two gameplay features to predict Normalized Learning Gain: ScannerCount and ReadingDuration. The model achieves a leave-one-out cross validation $R^2$ of 0.268. The negative coefficient of ScannerCount indicates that the more often students scan items in the virtual laboratory, the lower Normalized Learning Gain they achieve. Conversely, the longer students read, the higher their Normalized Learning Gain. This indicates that Presence is more challenging to predict from Gameplay features than Normalized Learning Gain.

### 5.2 Affect-Enhanced Composite Models

Next, we explored the potential of affect-enhanced predictive models for Normalized Learning Gain and Presence by introducing the 9 composite affective measures with the previous 8 gameplay features denoted Affect-Enhanced Composite models. From this set of 17 features, we performed the same forward stepwise feature selection technique optimizing for LOOCV $R^2$ and report the resulting models.

### Table 4. Resulting linear model for predicting Presence from gameplay features.

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>SE</th>
<th>$\beta$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversation Count</td>
<td>-31.9</td>
<td>17.5</td>
<td>-0.294</td>
<td>0.078</td>
</tr>
<tr>
<td>Worksheet Count</td>
<td>-52.6</td>
<td>23.6</td>
<td>-0.359</td>
<td>0.034*</td>
</tr>
</tbody>
</table>

Leave-One Out CV $R^2 = 0.147$

Note: ** - $p < 0.01$, * - $p < 0.05$

### Table 5. Resulting linear model for predicting Normalized Learning Gain from gameplay and composite-affect features.

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>SE</th>
<th>$\beta$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scanner Count</td>
<td>-0.369</td>
<td>0.237</td>
<td>-0.261</td>
<td>0.130</td>
</tr>
<tr>
<td>Reading Duration</td>
<td>0.0212</td>
<td>0.0106</td>
<td>0.364</td>
<td>0.049*</td>
</tr>
<tr>
<td>Surprise</td>
<td>-0.0183</td>
<td>0.0112</td>
<td>-0.246</td>
<td>0.111</td>
</tr>
</tbody>
</table>

Leave-One Out CV $R^2 = 0.324$

Note: ** - $p < 0.01$, * - $p < 0.05$

The optimal Affect-Enhanced Composite model for predicting Normalized Learning Gain uses three features and achieves a LOOCV $R^2$ of 0.324. Of the three features selected from a potential 17, two are gameplay features (ScannerCount and ReadingDuration), and one is a Composite feature (Surprise). It is important to note that these two gameplay features are the same
for the Baseline model for Normalized Learning Gain, indicating that adding Surprise to that model improves the predictive accuracy for Normalized Learning Gain from an LOOCV $R^2$ of 0.268 to 0.324. The coefficients of the two gameplay features are also similar to those of the Baseline model, indicating that the incorporation of Surprise improves the former model without making drastic changes to the former model parameter estimates. The negative coefficient of Surprise suggests that, holding other variables constant, more surprise leads to reduced Normalized Learning Gain.

### Table 6. Resulting linear model for predicting Presence from gameplay and composite-affect features.

<table>
<thead>
<tr>
<th>Affect-Enhanced Composite Model for Presence</th>
<th>B</th>
<th>SE B</th>
<th>$\beta$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversation Count</td>
<td>-39.3</td>
<td>16.4</td>
<td>-0.362</td>
<td>0.024*</td>
</tr>
<tr>
<td>Worksheet Count</td>
<td>-64.0</td>
<td>21.6</td>
<td>-0.437</td>
<td>0.006**</td>
</tr>
<tr>
<td>Disgust</td>
<td>-4.56</td>
<td>1.95</td>
<td>-0.369</td>
<td>0.027*</td>
</tr>
<tr>
<td>Confusion</td>
<td>-2.20</td>
<td>1.29</td>
<td>-0.325</td>
<td>0.099</td>
</tr>
<tr>
<td>Sadness</td>
<td>3.44</td>
<td>1.20</td>
<td>0.560</td>
<td>0.008**</td>
</tr>
</tbody>
</table>

Leave-One-Out CV $R^2 = 0.245$

Note: ** - p < 0.01, * - p < 0.05

The optimal Affect-Enhanced Composite model for Presence achieves a LOOCV $R^2$ of 0.245, improving on the Baseline LOOCV $R^2$ of 0.147. This model uses five features, with the same two gameplay features (Conversation Count and Worksheet Count) as the Baseline model and three features from the set of Composite measures (Disgust, Confusion, and Sadness). The negative coefficients of Disgust and Confusion indicate that students experiencing more duration with elevated standardized evidence score for these composite emotions exhibited lower Presence. Conversely, students with a higher proportion of expressed Sadness reported higher Presence. This Affect-Enhanced Composite model also indicates an additive effect in the sense that the same gameplay features were maintained, but the predictive accuracy is improved by incorporating several Composite features.

### 5.3 Affect-Enhanced AU Models

While the Affect-Enhanced Composite models improve upon the Baseline models, the Action Units (AUs) measured by iMotions provide low-level feature evidence scores that offer a more granular level of facial expression features than the Composite measures. Creating predictive models from this more granular measure of affect may produce a more accurate affect-enhanced model at the cost of interpretability since the AUs constitute a lower-level facial expression feature than the Composite measures. There are 20 AUs measured by iMotions, yielding a total feature set of 28 when including the gameplay features that can potentially be used in modeling students’ normalized learning gain and presence.

Using the forward stepwise feature selection optimizing LOOCV $R^2$ for predicting Normalized Learning Gain from the feature set of 20 AUs and 8 Gameplay features generates a model with six features and a LOOCV $R^2$ of 0.475. The two gameplay features selected (Scanner Count and Reading Duration) are the same as the Affect-Enhanced Composite model and Baseline model for Normalized Learning Gain. The other four features were AUs, including AU1 Inner Brow Raiser, AU4 Dimpler, AU15 Lip Corner Depressor, and AU43 Eyes Closed. Because the Sadness measure is partially composed of AU1 Inner Brow Raiser (as indicated by Table 1), the inclusion of AU1 Inner Brow Raiser shows the similarity between the affect-enhanced models because Sadness was included in the Composite-Affect Normalized Learning Gain model. The coefficients for both Gameplay features remain similar to both the Affect-Enhanced Composite Normalized Learning Gain model and the Baseline Normalized Learning Gain model. The addition of the AUs into this linear model provides a greater LOOCV $R^2$ than either the baseline Normalized Learning Gain model or Affect-Enhanced Composite Normalized Learning Gain model. Thus, the Affect-Enhanced AU Normalized Learning Gain model exemplifies the additive value of more granular affective measures in the predictive models for Normalized Learning Gain, compared to the Affect-Enhanced Composite Normalized Learning Gain model.

### Table 7. Resulting linear model for predicting Normalized Learning Gain.

<table>
<thead>
<tr>
<th>Affect-Enhanced AU Model for Normalized Learning Gain</th>
<th>B</th>
<th>SE B</th>
<th>$\beta$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scanner Count</td>
<td>-0.294</td>
<td>0.203</td>
<td>-0.207</td>
<td>0.159</td>
</tr>
<tr>
<td>Reading Duration</td>
<td>0.0353</td>
<td>0.0087</td>
<td>0.590</td>
<td>&lt; 0.001**</td>
</tr>
<tr>
<td>AU1</td>
<td>-0.0290</td>
<td>0.0085</td>
<td>-0.413</td>
<td>0.002**</td>
</tr>
<tr>
<td>AU14</td>
<td>-0.0548</td>
<td>0.0157</td>
<td>-0.534</td>
<td>0.002**</td>
</tr>
<tr>
<td>AU15</td>
<td>0.0470</td>
<td>0.0186</td>
<td>0.390</td>
<td>0.018*</td>
</tr>
<tr>
<td>AU43</td>
<td>-0.0340</td>
<td>0.0163</td>
<td>-0.265</td>
<td>0.048*</td>
</tr>
</tbody>
</table>

Leave-One-Out CV $R^2 = 0.475$

Note: ** - p < 0.01, * - p < 0.05

An Affect-Enhanced AU Presence model was created using the feature set of 20 AUs and 8 gameplay features and the same feature selection techniques as the previous models. This resulted in five selected features, the same two gameplay features from the previous Presence models (Conversation Count and Worksheet Count) along with three AUs. The included AUs were AU7 Lid Tightener, AU12 Lip Corner Puller, and AU28 Lip Suck. While the gameplay features are the same as the Affect-Enhanced Composite and Baseline Presence models, none of the AUs included in the Affect-Enhanced AU Presence model contribute to the composite features included in the Affect-Enhanced Composite model. The Affect-Enhanced AU Presence model results in an increased LOOCV $R^2$ over both the Affect-Enhanced Composite model and Baseline model. Since this model preserves similar coefficients for the same features used in the Baseline Presence model, it appears the AUs improve the predictive accuracy of the Baseline in terms of LOOCV $R^2$. 


The predictive accuracy of the three families of student models is shown in Figure 3. The results show that including facial expression features improved the predictive power of the models for both learning gains and presence. Similarly, models utilizing action unit features achieved higher predictive accuracy than models using the composite emotion labels. This superior performance was achieved for both learning gains and presence. While the models of learning gains outperform their counterparts for presence for each model structure, the Affect-Enhanced Composite Presence model shows a larger increase from the Baseline than the Affect-Enhanced AU Normalized Learning Gain model, while the Affect-Enhanced AU Normalized Learning Gain model shows a larger increase from utilizing AU features than the Presence model.

The Affect-Enhanced models use the same gameplay features as the Baseline models, with the addition of several affective features. Thus, the Baseline models are nested within the Affect-Enhanced models, with the Affect-Enhanced models acting as “full models” and the Baselines as “reduced models.” Since the Baseline models are nested within their respective Affect-Enhanced Composite and Affect-Enhanced AU models, an F-test provides a measure of the contributions of the additional features included in the full models over the reduced models. Results from this test show that the Affect-Enhanced Composite Normalized Learning Gain model’s addition of one additional feature (Surprise) does not show significant improvement over the Baseline Normalized Learning Gain model (F(1,29) = 2.70, p = 0.11). The Affect-Enhanced AU Normalized Learning Gain model’s addition of four features shows significant improvement over the Baseline Normalized Learning Gain model (F(4, 26) = 5.12, p = 0.004) in terms of reduction in residual sum of squares.

For the Presence models, results indicate that both the Affect-Enhanced Composite (F(3, 27) = 3.37, p = 0.033) and Affect-Enhanced AU (F(3, 27) = 3.95, p = 0.019) Presence models’ incorporation of three affective features show significant improvement over the Baseline Presence model in terms of reduction in sum of squared errors.

6 DISCUSSION

The results of the study suggest that affect-enhanced student models can improve upon predictive accuracy compared to models based exclusively on student behavior in the learning environment. By augmenting gameplay data with facial expression data, affect-enhanced student models significantly outperformed models using only gameplay features. Additionally, the models using the more granular action unit data significantly outperformed models using the composite emotions. These results highlight the importance of decisions about the granularity of affect data representation.

Across multiple feature sets, predictive models for Normalized Learning Gain outperformed models for Presence, as measured by leave-one-out cross validation R². In all of the Normalized Learning Gain models, ReadingDuration is included with high significance. The virtual books and articles contain information relevant to questions presented in the microbiology assessment, making it a strong candidate for predicting the learning gain assessed by the microbiology test. For both learning

<table>
<thead>
<tr>
<th>Conversation Count</th>
<th>B</th>
<th>SE B</th>
<th>β</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU7</td>
<td>4.89</td>
<td>1.58</td>
<td>-0.581</td>
<td>0.005**</td>
</tr>
<tr>
<td>AU12</td>
<td>3.88</td>
<td>1.61</td>
<td>0.440</td>
<td>0.023*</td>
</tr>
<tr>
<td>AU28</td>
<td>-2.01</td>
<td>0.822</td>
<td>-0.367</td>
<td>0.022*</td>
</tr>
</tbody>
</table>

Leave-One Out CV $R^2 = 0.296$

Note: ** - p < 0.01, * - p < 0.05
gains and presence, the models utilizing the more granular AUs achieved significantly higher predictive accuracy than those utilizing composite emotions. This could be due to a variety of factors. First, the automatic tagging of AUs may be more accurate, which is plausible as action units represent more concretely represented phenomena than the composite emotions. In addition, the facial expression tracking framework underlying the emotion recognition system was trained on a library of posed faces, which are potentially more expressive than emotions that appear in a learning context. It could also be the case that important information is being captured by AUs that is not captured by the composite emotion labels. For example, an AU may be conveying a subtle expression of an emotion not captured by the composite labeling. It is also possible that the AUs are representing a contextually-specific response not captured by the composite labels, such as signaling increased cognitive load.

Analysis of the specific emotions present in the composite models show that less Surprise is more predictive of learning gains. The result highlights the need to further investigate where in the gameplay activity the periods of Surprise tend to occur. Similarly, for presence measures, the model shows less Disgust and less Contempt and increased Sadness to be predictive of presence. While the relationship for Disgust and Contempt align with what one would expect for increased presence, increased Sadness does not appear to align, though could potentially be related to students’ empathy for the sick patients in the mystery or emotional investment in solving the mystery.

Many of the AUs found to be significant in this work have previously been found as commonly occurring AUs that have predictive significance for tutoring outcomes [23]. The presence of AU14 Dimpler being negatively associated with learning is a result also found in a study of affect in tutoring [24], which found AU14 Dimpler to be an indicator of lower engagement and higher frustration. AU1 Inner Brow Raise is associated with surprise, which was the only emotion present in the composite model, further strengthening the need to further investigate which game events are triggering the reaction. For the presence models, AU7 Lid Tightener and AU28 Lip Suck were negatively associated with presence. These AUs have been associated with increased mental effort, which may negatively impact student perceptions of “transportation” into the virtual environment.

While AUs may produce better models, using them sacrifices some of the interpretability found in the composite models. For example, one can imagine designing an adaptive intervention to induce confusion, or perhaps reduce frustration, but it is not as clear what the design implications would be for inducing or decreasing a specific AU. While some AUs can be related to composite emotions, others may be more difficult to categorize or require additional information to contextualize. On the other hand, composite emotions may not always be more interpretable, as shown by the features in the presence model. The trade-off between predictive ability and interpretability is one that requires further investigation, and may depend on the types of adaptivity that a game-based learning environment is designed to provide. One approach to improving interpretability would be to investigate how these models differ as students move through a game. While the data presented here was accumulated through complete playthroughs of a game, examining how facial expressions evolve as students transition between affective states throughout different phases of the gameplay session may help identify specific game events or contexts that produce the most desirable reactions. For example, expressions of frustration may be distributed evenly through the session, or they may be triggered by specific events. A more granular analysis of the data may help identify such scenarios, allowing for targeted and effective supports to be incorporated into the game-based learning environment.

7 CONCLUSION

Affect is integral to learning and problem solving in adaptive learning environments. Creating affect-enhanced student models has the potential to provide the foundation for the next generation of adaptive learning environments that are more effective and more engaging. Because game-based learning environments are specifically designed for the dual goals of increasing learning effectiveness and increasing student engagement, the prospect of introducing affect-enhanced student models into game-based learning environments may be a particularly fruitful line of investigation. Creating affect-enhanced student models that leverage facial expression analysis could potentially yield student models that can more accurately predict student learning and engagement and, ultimately, contribute to improved student experiences through more informed adaptation.

In this work we have investigated affect-enhanced student models with a game-based learning environment. Baseline models using gameplay features were created to predict normalized learning gain and presence using a forward stepwise feature selection methodology. Two sets of affect-enhanced student models, one based on composite emotion models and one based on action units, were created. Both affect-enhanced student models augmented gameplay data with these features sets. Results show that the affect-enhanced student models improve upon the predictive accuracy of baseline models. The additional features incorporated in the affect-enhanced models were also found to significantly contribute to predictive accuracy, suggesting that the affective features provide additive value. The action-unit-based affect-enhanced models performed better than the composite-based affect-enhanced models, achieving predictive accuracy improvements for both student learning (as measured by normalized learning gain) and student engagement (as measured by presence). In future work, it will be important to investigate the integration of affect-enhanced student models into runtime game-based learning environments. Explorations of runtime integration should include early prediction modeling and game-based adaptation that uses the affect-enhanced student models to guide both cognitive and affective scaffolding and to understand their impact on learning processes and outcomes.

Acknowledgements

We thank our collaborators in the Center of Educational Informatics and the SMART Lab at North Carolina State University. This study was supported by funding from the Social Sciences and Humanities Research Council of Canada. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Social Sciences and Humanities Research Council of Canada.
REFERENCES


