British Journal of Educational Technology *Vol 0 No 0 2020* 1–22
doi:10.1111/bjet.12992

Multimodal learning analytics for game-based learning

Andrew Emerson, Elizabeth B. Cloude , Roger Azevedo and James Lester

Andrew Emerson is a doctoral student in Computer Science at North Carolina State University. His research spans multimodal learning analytics, predictive student modeling and educational applications of machine learning. He received a BS in Computer Science and Mathematics from Furman University. Elizabeth B. Cloude is a graduate student in Learning Sciences and Educational Research at the University of Central Florida. Her research is focused on capturing the complexity of human learning using multimodal data to design educational systems that capture, monitor and adapt to meet individual learning needs to optimize performance. She received a BS in Biopsychology at Christopher Newport University. Roger Azevedo is a Professor of Learning Sciences and Educational Research at the University of Central Florida. His primary research includes examining the role of cognitive, metacognitive, affective and motivational self-regulatory processes during learning with advanced learning technologies. He is a Fellow of the American Psychological Association (APA). James Lester is a Distinguished University Professor of Computer Science and Director of the Center for Educational Informatics at North Carolina State University. His research focuses on artificial intelligence technologies for education. He is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI). Address for correspondence: Andrew Emerson, Department of Computer Science, North Carolina State University, Engineering Building III, Raleigh, NC 27695-8206, USA. Email: ajemerso@ncsu.edu

Abstract

A distinctive feature of game-based learning environments is their capacity to create learning experiences that are both effective and engaging. Recent advances in sensor-based technologies such as facial expression analysis and gaze tracking have introduced the opportunity to leverage multimodal data streams for learning analytics. Learning analytics informed by multimodal data captured during students' interactions with game-based learning environments hold significant promise for developing a deeper understanding of game-based learning, designing game-based learning environments to detect maladaptive behaviors and informing adaptive scaffolding to support individualized learning. This paper introduces a multimodal learning analytics approach that incorporates student gameplay, eye tracking and facial expression data to predict student posttest performance and interest after interacting with a game-based learning environment, Crystal Island. We investigated the degree to which separate and combined modalities (ie, gameplay, facial expressions of emotions and eye gaze) captured from students ($n = 65$) were predictive of student posttest performance and interest after interacting with Crystal Island. Results indicate that when predicting student posttest performance and interest, models utilizing multimodal data either perform equally well or outperform models utilizing unimodal data. We discuss the synergistic effects of combining modalities for predicting both student interest and posttest performance. The findings suggest that multimodal learning analytics can accurately predict students' posttest performance and interest during game-based learning and hold significant potential for guiding real-time adaptive scaffolding.

Practitioner Notes

What is already known about this topic

- Game-based learning has been shown to be effective for a broad array of subject matters and student populations.
- Sensor-based technologies are becoming increasingly powerful and cost-effective and hold significant potential for learning technology.
- Multimodal learning analytics are showing promise for their potential to provide insight into student learning.

What this paper adds

- Investigates a multimodal learning analytics approach for prediction of performance and interest in game-based learning.
- Demonstrates the joint effect of student gameplay and gaze data for predicting student interest.
- Demonstrates the synergistic effects of student gameplay and facial expression data for predicting student posttest performance and explains how this combination of modalities can inform real-time scaffolding.

Implications for practice and/or policy

- Eye tracking and facial expression tracking can provide insight into student learning processes in game-based learning, which can drive the feedback adaptive game-based learning environments provide to students.
- Researchers and practitioners should consider privacy and ethical concerns in storing multimodal data (eg, storing facial movements only for the extent of the interaction).
- Researchers and practitioners should determine the most appropriate representations of multimodal data that suit their learners (eg, age, expertise level), tasks (eg, well- vs. ill-structured) and domains (eg, biology, math).
- Researchers and practitioners should consider both the correlations between modalities when constructing predictive models and the diversity of the student populations from which these data are generated to ensure algorithmic bias is mitigated.

Introduction

Predictive student modeling holds significant promise for improving learning experiences. By observing students' learning activities over time, models could predict student knowledge and interest, which could then be used in real-time to inform adaptive scaffolding. Prior research has investigated student models of knowledge (Min *et al.*, 2017; Shute, Wang, Greiff, Zhao, & Moore, 2016; Spires, Rowe, Mott, & Lester, 2011) affect (Baker, D'Mello, Rodrigo, & Graesser, 2010; Botelho, Baker, & Heffernan, 2017; Henderson, Rowe, Mott, *et al.*, 2019). Several studies have demonstrated the value of incorporating multiple modalities of data generated from sensor-based technologies (eg, facial expressions of emotions, eye gaze) to model students' knowledge of complex topics (Blikstein & Worsley, 2016; Sharma, Papamitsiou, & Giannakos, 2019; Taub, Sawyer, Smith, *et al.*, 2020).

Recent years have seen a growing interest in multimodal learning analytics (Azevedo & Gašević, 2019; Blikstein, 2013; Blikstein & Worsley, 2016; Ochoa & Worsley, 2016; Oviatt, Grafsgaard, Chen, & Ochoa, 2018). Multimodal data have been widely used in machine learning (Baltrušaitis, Ahuja, & Morency, 2018) and it can be leveraged to understand students' learning behaviors at high granularities (eg, millisecond level of onset of disinterest; Aslan *et al.*, 2019). Data channels

used to model students' knowledge include learner-system interaction behavior and physiological measurements (Lane & D'Mello, 2019). Physiological sensors such as electroencephalography (EEG) and electrodermal activity (EDA) in combination with eye gaze and video recordings of facial expressions of emotions have shown significant promise in predicting learning outcomes when captured during learning with intelligent tutoring systems and other learning technologies (Blikstein & Worsley, 2016; Giannakos, Sharma, Pappas, Kostakos, & Velloso, 2019; Lane & D'Mello, 2019; Sharma, Papamitsiou, Olsen, & Giannakos, 2020; Vrzakova, Amon, Stewart, Duran, & D'Mello, 2020). Results have shown how physiological data, including eye gaze and facial expression, explain students' knowledge, attention, affect, learning strategies and other learning-related outcomes. They have also shown how multimodal data have the potential to improve predictive modeling with implications for improving pedagogical intervention and adaptation that support cognitive, affective and metacognitive aspects of learning. As such, it is critical to model relationships using multimodal data between students' posttest performance and characteristics related to learning (eg, interest) to gain insight into designing adaptive learning technologies that tailor instructional experiences to meet the individual learning needs of students (Di Mitri, Schneider, Specht, & Drachsler, 2018).

Multimodal learning analytics hold significant promise for game-based learning (Plass, Mayer, & Homer, 2020). However, few studies have examined relationships between unimodal and multimodal data to understand the joint effect of separate modalities and their capacity to explain students' posttest performance and constructs related to learning (eg, interest) when engaged in game-based learning. Multimodal learning analytics pose significant challenges that call for further study and analysis (Azevedo, Taub, & Mudrick, 2018). In some applications, such as affect detection during learning, multimodal data channels can have superadditive effects on the accuracy of predictive models for affective states, but sometimes multimodal data channels can have effects that are redundant or even inhibitory (D'Mello & Graesser, 2010). This calls for developing a better understanding of the multimodal data channels for learning analytics. In this paper, we address the aforementioned gaps by exploring relationships between students' prior knowledge, eye gaze, facial expressions of emotions and gameplay behavior traces to posttest performance and self-reported interest after game-based learning.

Game-based learning environments integrate game content with learning activities (eg, exploring, navigating, investigating) to enhance domain-specific knowledge (eg, microbiology) and skill acquisition (eg, self-regulation), where activities typically involve problem solving and challenges to foster students' perceived achievement (eg, solve a mysterious illness outbreak). These environments incorporate storylines with visual aesthetics that have been shown to motivate students to stay engaged during learning (Plass, Homer, & Kinzer, 2015). They often include incentive structures (eg, providing a treatment solution to a disease outbreak on a tropical island) and game mechanics that students find interesting (eg, solving a mysterious illness as a Center for Disease Control agent; Rotgans & Schmidt, 2011). Empirical studies of game-based learning have found that students achieve higher learning gains relative to traditional learning environments such as classroom settings (Mayer, 2019; Vlachopoulos & Makri, 2017). Additional studies have found that students report higher motivation and interest toward learning content with these environments relative to traditional settings (Qian & Clark, 2016).

A critical component of game-based learning is interest (Plass *et al.*, 2020). Motivational frameworks, such as self-determination and intrinsic motivation theories, describe interest as being related to perceived control over learning activities (ie, agency) and incentives (eg, rewards; Mayer, 2019; Wentzel & Miele, 2016). Ryan and Deci's (2000) motivation framework on intrinsic motivation describes interest as a result of satisfying needs for autonomy and competence

when completing learning activities, often leading to higher performance outcomes (Ryan & Deci, 2000). Since interest plays a crucial role in whether or not students' acquire knowledge, it is critical to capture and examine students' interest in relation to performance after game-based learning.

The majority of studies assessing relationships between students' interest and performance use traditional methodological techniques such as administering self-report measures *before* and *after* game-based learning (Plass *et al.*, 2020), missing critical information on students' interest *during* learning with game-based learning environments. Solely relying on self-report data administered before and after game-based learning is a major methodological and analytical issue (Ainley & Ainley, 2019; du Boulay & Del Soldato, 2016). We argue that predicting students' interest based on multimodal data generated during game-based learning will provide insight into capturing student interest in realtime and also provide a means for understanding relationships between students' interest and knowledge acquisition *during* game-based learning.

Related work using multimodal data for game-based learning

A number of studies have examined knowledge acquisition with game-based learning environments using a range of data channels such as gameplay behavior traces (eg, Alonso-Fernández *et al.*, 2019; Alonso-Fernández, Martínez-Ortiz, Caballero, Freire, & Fernández-Manjón, 2020; Taub *et al.*, 2017), facial expressions of emotions (eg, Lane & D'Mello, 2019; Taub, Sawyer, Smith, *et al.*, 2020), performance measures (eg, pre/posttest scores; Dever & Azevedo, 2019), self-report measures (eg, Cloude, Taub, Lester, & Azevedo, 2019) and eye gaze (eg, Dever, Wiedbusch, & Azevedo, 2019; Gomes, Yassine, Worsley, & Blikstein, 2013; Lee, Donkers, Jarodzka, & van Merriënboer, 2019; Tsai, Huang, Hou, Hsu, & Chiou, 2016). A noteworthy study used gameplay behavior traces to model students' self-reported interest during game-based learning (Sawyer, Rowe, Azevedo, & Lester, 2018). Students' gameplay behavior traces were operationalized as time spent interacting with various elements within the game such as the virtual books that introduce subject matter and the non-player characters where variables were calculated to account for relative time spent learning in the game. Their model highlighted how various gameplay behavior features may be more predictive of student interest in game-based learning depending on the learning context (eg, game-based learning in a classroom vs. a laboratory), indicating more studies are needed to derive more general conclusions.

Similarly, other studies have used facial expressions of emotions to examine its relation to students' knowledge acquisition since studies have found that emotions play a critical role in learning complex information (D'Mello & Graesser, 2010; D'Mello, Dieterle, & Duckworth, 2017; Lane & D'Mello, 2019; Loderer, Pekrun, & Lester, 2018). A noteworthy study contextualized students' facial expressions of emotions by operationalizing these data in relation to their expression during gameplay (Taub, Sawyer, Lester, *et al.*, 2020). Specifically, emotions were defined based on whether they were expressed during gameplay actions that were relevant to the overall objective of the learning environment (ie, how relevant were gameplay behavior traces to solving the mysterious illness during game-based learning). Results demonstrated that the context in which specific emotions were expressed (ie, emotions expressed during relevant vs. irrelevant gameplay actions) were related to knowledge acquisition after game-based learning. Their findings highlight the critical importance of modeling facial expressions of emotions to examine their relationship to knowledge acquisition (Taub, Sawyer, Lester, *et al.*, 2020). A study examining facial expressions of emotions also found that these data were related to students' interest during learning (Hidi & Renninger, 2019). Yet, few studies have used multimodal data such as facial expressions of emotions (Hidi & Renninger, 2019) in conjunction with gameplay behavior traces (Sawyer *et al.*, 2018) to explain students' self-reported interest.

Other data channels used to model knowledge include eye-gaze data (D'Mello *et al.*, 2017; D'Mello, Olney, Williams, & Hays, 2012). Studies reveal that patterns of eye movements in combination with gameplay behavior traces were indicative of how students learned during game-based learning environments (Taub *et al.*, 2017). Results showed that students who read virtual books in-depth (ie, rereading a book or spending longer times reading a book) during game-based learning demonstrated higher knowledge acquisition relative to students who did not read virtual books in-depth after game-based learning. Further, eye-gaze data have also been tied to motivation and emotions (Lallé, Conati, & Azevedo, 2018), what students' are attending to during game-based learning (Hutt *et al.*, 2019) and performance measures (Rajendran, Kumar, Carter, Levin, & Biswas, 2018). Additional studies have found that low and high performing students demonstrate differences in their patterns of eye movements during game-based learning (Gomes *et al.*, 2013). This study showed that the lengths of fixations on relevant learning material at key moments were indicative of problem-solving strategies or lack thereof. As such, representing granular traces of students' gameplay behaviors have shown to contribute to a deeper understanding of gameplay or learning processes, during game-based learning that contribute to knowledge acquisition.

In sum, literature has shown that multimodal data can paint a rich picture of students' learning with game-based learning environments by modeling relationships between multimodal data and knowledge acquisition (Taub *et al.*, 2017) and affective states (Henderson, Emerson, Rowe, & Lester, 2019). Henderson, Emerson, *et al.* (2019) used student posture and galvanic skin response to model affective states during learning, showing improved performance with both modalities in comparison to unimodal models. However, significant gaps remain in literature using multimodal data as few studies have explored the degree to which three commonly used data channels (ie, gameplay behavior traces, facial expressions of emotions and eye gaze), separately and/or jointly, can accurately predict student knowledge and interest. For instance, what is the ideal combination of data channels to predict knowledge acquisition with a game-based learning environment and is this combination the same when predicting interest? Little work has examined the joint effect of individual multimodal streams for predicting both posttest performance and no work has explicitly utilized multimodal data to model students' interest. Understanding the role of multimodal data and its relation to knowledge acquisition and interest have implications for designing adaptive and personalized game-based learning environments that have the capability to detect students' developing competency and interest using their multimodal data captured *during* learning activities and effectively intervene and scaffold students' when they demonstrate maladaptive behaviors detrimental to performance and interest.

Research objectives and research questions

While recent studies have shown that multimodal data can perform well in predictive models for student performance (Taub *et al.*, 2017), little work has examined which combination of modalities most accurately predict students' posttest scores and few studies have explicitly modeled student interest in game-based learning using a multimodal perspective. We address gaps in literature by examining relationships between students' level of prior knowledge (ie, pretest scores), eye gaze, gameplay behavior traces, facial expressions of emotions, performance outcomes and self-reported interest. Specifically, we built predictive models to assess relationships between students' knowledge acquisition (ie, posttest scores) and self-reported interest after interacting with Crystal Island by combining multimodal data streams. By adopting a multimodal perspective, we investigated the degree to which separate and combined modalities of data (eg, gameplay, facial expression of emotions and eye gaze) explain knowledge acquisition and self-reported interest after game-based learning with Crystal Island. Our research questions are below:

RQ1: How well do combinations of student gameplay behavior traces, facial expressions of emotions and eye gaze classify low, medium and high performing groups of students after game-based learning?

RQ2: How well do combinations of student gameplay behavior traces, facial expressions of emotions and eye gaze classify low, medium and high interest groups of students after game-based learning?

Methods

Participants and Materials

College students were recruited from three large North American universities and interacted with Crystal Island in a controlled, laboratory setting. Participants were compensated $10/hour for completing the study, as described in Taub, Sawyer, Smith, *et al*. (2020). IRB approved the study prior to recruiting where the study posed minimal to no risks for participating. For this study, participants were randomly assigned to one of three conditions where each were designed with varying levels of agency. For this paper, we only used participants in the *full agency* condition because students were afforded complete control over their actions during game-based learning, such that they could select and interact with all elements at any point without restrictions. We do not describe the other experimental conditions because these participants were not included in the analyses. Additionally, we did not want the design of experimental manipulations to influence multimodal data generated during game-based learning. In this condition, there were sixty-five ($n = 65$) college student participants. Four students were eliminated from our data set because critical survey or sensor data were missing. This resulted in a sample of 61 students ($M = 20.1$ years old, $SD = 1.56$) of which 42 ($n = 42$; 69%) were female. Each student played the game until correctly solving the mystery or ran out of time (maximum of 3 hours). Gameplay durations ranged from 26.5 to 159.9 minutes ($M = 68.2$, $SD = 22.7$).

Prior to learning with Crystal Island, participants completed a series of questionnaires as well as a 21-item, 4-option multiple pretest assessment to measure prior knowledge about microbiology ($M = 11.84$, $SD = 2.74$). Out of the 21 items, there were 12 factual questions (eg, *What is the smallest type of living organism?*) and nine application questions (eg, *What is the difference between bacterial and viral reproduction?*). After interacting with Crystal Island, the Intrinsic Motivation Inventory (IMI; Ryan, 1982) was administered, a validated survey consisting of a 7-point Likert scale (1 = Not true at all, 4 = Somewhat true, 7 = Very true) with 29 items and five subscales related to intrinsic motivation (McAuley, Duncan, & Tammen, 1989). For purposes of this study, the Interest-Enjoyment subscale ($\alpha = 0.96$; $M = 4.67$, $SD = 1.37$) was the primary subscale utilized, which consisted of seven items: (1) I enjoyed doing this activity very much; (2) This activity was fun to do; (3) I thought that this was a boring activity; (4) This activity did not hold my attention at all; (5) I would describe this activity as very enjoyable; (6) I thought this activity was quite enjoyable; and, (7) While doing this activity, I was thinking about how much I enjoyed it. We do not provide information on the other IMI subscales due to space limitations and because those data were not used in the analyses. Afterward, participants completed a 21-item, 4-option multiple-choice posttest assessment similar to the pretest assessment to capture acquired knowledge about microbiology ($M = 14.13$, $SD = 2.85$). The pretest and posttest assessment sequence of items were randomized from pretest to posttest administration to reduce practice effects. Both the pretest and posttest are validated instruments that were adopted from Nietfeld, Shores, and Hoffmann (2014). Questionnaires not included in our analyses were not included due to space limitations. See Taub, Sawyer, Smith, et al. (2020) for specific details.

Crystal Island: A game-based learning environment

Crystal Island is a game-based learning environment for microbiology education (see Figure 1; Rowe, Shores, Mott, & Lester, 2011). Students take on the role of a medical researcher who

has arrived on a remote island research station. When students arrive at the island, they discover a mysterious disease outbreak plaguing the members of the research staff. Crystal Island is an open-world 3D game, giving students the ability to freely explore within the environment. Student gameplay actions, including movement, dialogue and interactions with in-game objects (eg, books, scanner, and food) are all recorded in gameplay behavior traces that are used for subsequent data analysis. To solve the mystery, students must gather evidence and test hypotheses related to microbiology concepts surrounding the disease.

Experimental procedure

Participants entered the laboratory setting and were greeted by a researcher who instructed them to sit in front of a computer for a study that was conducted over a single session. Upon obtaining informed consent, the researchers randomly assigned participants to one of three conditions. Afterward, participants completed the pretest assessment prior to game-based learning with Crystal Island. The researchers then calibrated an eye tracker and facial expressions of emotions software for the participants (refer to Figure 2 of experimental set up). Once calibration was completed, participants were instructed to begin interacting with Crystal Island, which started with a tutorial that introduced participants to Crystal Island and the overall objective of the learning session, which was to solve the mystery illness plaguing the inhabitants on the island. Participants were then required to interact with Crystal Island for up to 180 minutes or complete the game by solving the science mystery. After gameplay, participants completed the Interest-Enjoyment subscale on the IMI questionnaire and the posttest capturing acquired knowledge of microbiology. Participants were then thanked, debriefed and compensated for their time in the study.

Coding and scoring

Student knowledge

To capture prior knowledge, we calculated the total number of correct answers on the pretest ($M = 11.84$, $SD = 2.74$) which were included as a feature for the predictive models. Similarly, we calculated the total number of correct answers on the posttest ($M = 14.13$, $SD = 2.85$; min = 8.0, max = 20.0) to operationalize student knowledge and included these data as a target variable in our predictive model). Specifically, we standardized the pretest score by subtracting each score by the mean and then, dividing it by the standard deviation. For the posttest target variable, we converted the predictive task into a classification problem by splitting the posttest scores into tertiles defined by the distribution of scores, where each group contained one-third of the sample scores.

Figure 1: The Crystal Island game-based learning environment

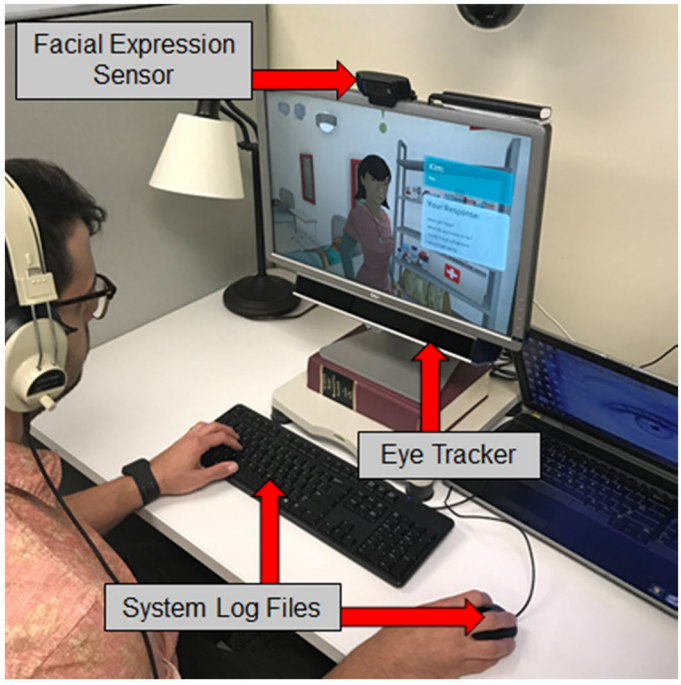

Figure 2: Fully instrumented participant and Crystal Island *environment*

Next, we assigned participants to either a low (less than a posttest score of 13.0; 21 students), medium (between a posttest score of 13.0 and 16.0; 21 students) or high (greater than a posttest score of 16.0; 19 students) posttest performance group. This approach follows prior work on creating a fine-grained, balanced classification problem for learning analytics (Akram *et al.*, 2018; Min *et al.*, 2020). We chose to split the data using this method over using the raw posttest values because of the limited sample size and we chose tertiles over a median split to achieve higher granularity.

Student interest

In addition to taking the posttest after interacting with Crystal Island, each student completed the Intrinsic Motivation Inventory (IMI; Ryan, 1982). To model interest, we used scores captured from the Interest and Enjoyment subscale ($M = 4.67$, $SD = 1.37$). Similar to the knowledge target variable grouping procedure, we converted interest and enjoyment scores to groups based on tertiles. The tertile approach simplifies the prediction task while maintaining some levels of granularity. For this variable, we split students into a low (less than or equal to a score of 4.14; 21 students), medium (between a score of 4.14 and 5.23; 21 students) or high (greater than a score of 5.23; 19 students) interest group. This approach creates a balanced distribution of classes and achieves a more fine-grained representation than a high versus low binary split of interest groups.

Facial expression recognition and feature representation

To study students' facial expressions of emotions, we equipped Crystal Island with a video-based facial expression tracking system, FACET (iMotions, 2016). This system extracts features that correspond to the Facial Action Coding System (FACS) (Ekman & Rosenberg, 1997). This software analyzes each video frame of each student's face and classifies 20 facial action units (AUs),

as well as nine composite emotional states (ie, *Sadness*, *Fear*, *Anger*, *Disgust*, *Contempt*, *Surprise*, *Frustration*, *Confusion* and *Joy*), defined as deviations from the baseline affective state established during calibration. FACET computes evidence scores for each AU and composite emotion, which represent the log odds of the presence of a facial expression as coded by a trained human coder. While we collected both individual AUs and composite emotions, this study only analyzed AUs data since they have been shown to predict learning in prior work (Sawyer *et al.*, 2017; Taub *et al.*, 2019).

We followed the same preprocessing steps used in Sawyer *et al.* (2017) to represent the facial expression features in a way that is suitable for predictive student modeling. First, we standardized the evidence scores of each AU data point for all participants by subtracting the evidence score by the mean and then, dividing it by the standard deviation over the individual student's entire gameplay experience. The mean and standard deviation for each AU are calculated per student, as this step supports a fair comparison of the relative change in facial expressions between students. Second, we filter the sequence of the standardized AU data to include only AU events where the evidence score rose above one standard deviation. This ensures that each AU event captured represents positive evidence of that particular AU. Third, we remove any AU events that did not occur for longer than a threshold of 0.5 seconds in order to remove the effect of microexpressions in the AU sequence. Finally, we sum the duration (in seconds) for each of the 20 AU events for each student to capture a static representation of the student's facial expression over their gameplay. Ultimately, the features used in the predictive student models are the total duration for each AU divided by the total time that each student spent playing Crystal Island. This representation encodes the proportion of gameplay duration for which each student exhibited positive evidence above one standard deviation for each AU. This higher level representation, in comparison to the original sequence of AU evidence scores, is a more compact encoding of student facial expressions and can be more easily utilized by standard machine learning models. While this representation loses the variability of AUs over time, it does enable the use of less computationally demanding modeling techniques that can explain relationships between the input and output data. In total, there are 20 facial expression features for each student. We will refer to models that utilize facial expression data and features derived from this source with the term "Face."

Gaze-based entity tracking and eye tracking feature representation

To incorporate eye gaze data into the predictive models, we captured eye gaze during gameplay with the SMI RED 250 eye tracker using a 9-point calibration. During each student's interaction with Crystal Island, the software responsible for logging student eye gaze data analyzes eye movements and finds the precise point on the computer screen where the student is looking. Eye movements were tracked at a rate of 120 Hz and logged with the corresponding timestamp. The eye gaze logging software then identified fixations using a standard minimal threshold of 250 milliseconds (Rayner, 1998). Within the Crystal Island software, we also implemented eye-gaze logging software that detected the in-game element (eg, book or research article) the student fixated upon by analyzing the angle and the gaze point on the screen. The logging software utilizes ray casting to find the intersection of the game object and the point on the computer screen in real-time, thus, allowing for a synchronized sequence of fixations. The final gameplay behavior trace has a record of the duration of the fixation and the name of the in-game object. To represent this data channel in a more compact encoding that is readily usable by standard machine learning models, we compute the total duration that the student spent fixating upon each in-game object. Within Crystal Island, there are 144 individual game objects which are specific to the game's narrative. We group these objects into eight broader categories based on their context within the game. The categories were chosen to group the in-game objects tracked by the eye-tracker as

higher level game-based learning objects. We did not explicitly contextualize the gaze categories based on phases of scientific reasoning processes or self-regulated learning as has been done in previous work (Taub *et al.*, 2017), but rather we categorize objects based on the context of the game itself using a bottom-up approach (Emerson, Sawyer, Azevedo, & Lester, 2018). The categories used in this work are listed in Table 1.

We selected these categories to create a high-level representation of student gaze patterns in relation to the game mechanics in Crystal Island and the differential diagnosis task students perform when interacting with it. After calculating the duration the student spent gazing upon objects in each category, we converted the durations into proportions of time for each individual category. In addition to the gaze categories, we incorporated a second metric for student gaze, fixations-per-second. This is calculated as the total number of fixations for each student, divided by their total game time. This feature representation has shown to be predictive of student engagement in prior work (Emerson *et al.*, 2018). The method of using broad categories is a step toward a generalized approach to modeling student gaze. For example, "Lab" and "Diagnosis" objects can be considered as a form of applying scientific reasoning processes in association with generating hypotheses. These patterns can help identify what a student is attending to during their gameplay. In total, there are nine eye gaze features for each student (8 object categories and fixations-per-second). We will refer to models that incorporate eye gaze features and features that are derived from this source with the term "Gaze."

Gameplay features

In addition to the sensor-based data collected by the eye-tracker and FACET software, gameplay behavior logging software in Crystal Island recorded each in-game action performed by the students. In these data, we distinguish eight action types. These action types include movement to a new location in the game, conversation with a character in the game, reading a book, completing an in-game achievement (eg, correctly solving the mystery), scanning a food item to test for a contaminant and recording findings in the diagnosis worksheet. We include student gameplay actions, which may reveal their interest, a key component of the multifaceted construct of emotional engagement (Fredricks, Blumenfeld, & Paris, 2004). Following previous work that incorporates gameplay features into student models, we created a one-hot encoding of each action the student performs in the Crystal Island game (Min *et al.*, 2017), such that each time a student performed one of these actions, we converted it to its corresponding one-hot encoding,

Table 1: Eye gaze object categories

Category	*Description*
Non-Player Characters (NPCs)	The set of virtual characters with which the student interacts
Setting	Objects the student encounters when traveling within the game
Food	Food items to test for contaminants
Lab	Equipment to test their hypothesis that are relevant to solving the mystery
Diagnosis	Notes about the student's findings
Book	Virtual material that students read throughout their gameplay to gather evidence, which involves the use of several learning strategies and self-regulated learning strategies
Concept Matrix	In-game assessments students complete after reading a virtual scientific book
Miscellaneous	Elements in the game that are not associated with game content, such as the heads-up display, settings menu and achievement panel

operationalized as a vector of size eight, populated with all zeros except for the element that maps to the specific action, which is represented by a one. After converting each student's game log into this format, we summed the one-hot encoded vectors over the entirety of the student's gameplay, yielding a sum of each action type. We then converted these sums to proportions of how often the student performed each action divided by how many actions he or she performed in total. This transformation process was inspired by previous work that has represented student actions from gameplay data (Geden, Emerson, Rowe, Azevedo, & Lester2020; Min *et al.*, 2017). We also include the total time the student spent playing Crystal Island as a feature. In total, there are nine gameplay features for each student (8 game action types and gameplay duration). We will refer to models that incorporate gameplay features and the features derived from this source with the term "Gameplay."

Multimodal predictive modeling

To investigate a multimodal perspective to predictive student modeling, we built predictive models of student performance and interest that were informed by multiple modalities of data: (1) gameplay behavior traces, (2) facial expressions of emotions and (3) eye gaze to classify both knowledge and interest separately into low, medium and high groups based on their scores. We compared different sets of multimodal predictive classifiers trained on a data set of the 61 students who learned with Crystal Island. The total number of features from all modalities includes 20 related to facial expression, nine related to eye gaze and nine related to gameplay, for a total of 38 features per student. The training and evaluation of each predictive model was conducted using 10-fold stratified cross-validation at the student-level, allowing all 61 students to be used for training and testing. We performed stratified cross-validation to maintain a training set distribution of the dependent variable that resembles the overall variable distribution. In practice, this meant that each training fold had approximately an equal number of each class represented in the data. This technique is critical with small data samples, such as ours with 61 students. A common issue when creating predictive models with relatively little data is having too many features, which causes models to overfit. To compensate, we performed a feature selection within each cross-validation fold to select the most informative features. The specific method we used selected the 15 most informative features when predicting the training set's dependent variable with the training data. The algorithm calculates the ANOVA F-value for each feature and then, selects those with the highest scoring features. If the specific data set has less than or equal to 15 features, we used all of the features provided. Within each cross-validation fold, we standardized the training and testing data by using the mean and standard deviation of the training data. We also ensured that there was no overlap of students in each training and testing set to avoid data leakage. For both prediction tasks, the specific machine learning model we employed was a logistic regression model with an L1 loss function. The L1 loss supports further feature selection by setting features that do not provide useful information to have coefficients equal to zero. We adopted the logistic regression model for its relative simplicity and interpretability. The metrics shown in the results are the best performing model configurations for each combination of multimodal data. The parameters used in the logistic regression models are discussed in the results section. All predictive models in this work were constructed in Python 3 using the Scikit-learn library (Pedregosa *et al.*, 2011).

Results

RQ1: How well do combinations of student gameplay behavior traces, facial expressions of emotions and eye gaze classify low, medium and high performing groups of students after game-based learning?

To examine how well combinations of students' gameplay behavior traces, facial expressions of emotions and eye gaze classified low, medium and high performing groups after game-based

learning, models were designed to predict students' posttest performance at each moment in time throughout gameplay sessions. In contrast to a learning gains approach, this approach aims to find relationships between posttest knowledge and variables from the multimodal data streams without controlling for prior knowledge. To predict whether a student was in the low, medium or high posttest performance class, we trained logistic regression models and compared each combination of modalities, determining the most effective multimodal data combinations. To allow for better tuning of the logistic regression models, we varied the regularization penalty parameter, C, between 0.25, 0.5 and 1.0, where a smaller value corresponds to higher regularization and fewer features with nonzero coefficients. To compare the results of the predictive models, we report the cross-validation accuracy of each model. The baseline to which we compare all models is the majority class prediction. Because the classes are almost equally distributed, the baseline performance is 34.4% based on the majority class of the tertile split. The results for the predictive models with different combinations of the multimodal data are shown in Table 2. The coefficients for the non-zero features of the best performing models, Gaze and Gameplay + Face, are shown in Tables 3 and 4, respectively.

RQ2: How well do combinations of students' gameplay behavior traces, facial expressions of emotions and eye gaze classify low, medium and high interest groups of students after game-based learning?
To examine how well combinations of students' gameplay behavior traces, facial expressions of emotions and eye gaze classified low, medium and high interest groups of students after game-based learning, we used the same machine learning pipeline to compare the most predictive combination of modalities. To tune the logistic regression models, we again varied the C parameter by 0.25, 0.5 and 1.0. We report the cross-validation accuracy of each model and we compare each model against a baseline of 34.4% due again to the tertile splits. The results for the predictive student models of interest can be seen in Table 5. The coefficients for the non-zero features of the best performing model, Gameplay + Gaze, are shown in Table 6.

Discussion

In this paper, we investigated the effectiveness of combining multimodal data streams to classify low, medium and high performance groups as well as low, medium and high interest groups after game-based learning. We investigated the degree to which unimodal and combined modalities (eg, gameplay, facial expression of emotions and eye gaze) classify performance and interest after interacting with Crystal Island. Our results suggested that multimodal predictive models can accurately classify both low, medium and high performance groups as well as low, medium and high interest groups.

Table 2: Posttest score prediction results (bold values represent best performance)

Data used	*Accuracy*
Majority class baseline	0.344
Gaze	**0.607**
Face	0.574
Gameplay	0.557
Gaze + Face	0.541
Gameplay + Gaze	0.574
Gameplay + Face	**0.607**
Gameplay + Gaze + Face	0.557

Table 3: Feature coefficients for the gaze posttest performance model

Feature	*Low*	*Medium*	*High*
Gaze-NPC	0.0	0.012	0.0
Gaze-Setting	0.149	−0.215	0.0
Gaze-Food	0.001	0.0	0.016
Gaze-Diagnosis	0.0	0.029	−0.017
Gaze-Miscellaneous	0.0	0.124	−0.203
Gaze-Book	−0.048	0.0	0.003
Gaze-ConceptMatrix	0.0	0.190	−0.013
Gaze-FixationsPerSecond	0.0	0.0	−0.016
Pretest Score	−0.758	0.283	0.118

Table 4: Feature coefficients for the gameplay + face posttest performance model

Feature	*Low*	*Medium*	*High*
Gameplay-Books	0.0	−0.032	0.0
Gameplay-Conversation	0.031	−0.130	0.0
Gameplay-Movement	0.0	−0.032	0.012
Gameplay-Posters	−0.157	0.184	0.0
Gameplay-Scanner	0.013	0.0	0.0
Gameplay-WorksheetSubmit	0.031	0.0	0.0
GameTime	0.0	−0.024	0.0
Face-AU15	0.201	−0.286	0.0
Face-AU23	0.014	0.0	0.0
Face-AU7	−0.028	0.0	0.004
Face-AU10	0.0	−0.079	0.0
Face-AU28	0.0	−0.123	0.141
Face-AU14	0.197	−0.117	0.0
Face-AU24	0.010	−0.014	0.0
Face-AU12	0.004	0.0	0.0
Face-AU4	0.0	0.147	−0.005
Face-AU25	0.0	−0.046	0.0
Face-AU20	0.157	0.0	0.0
Face-AU17	0.001	−0.024	0.0
Face-AU43	−0.146	0.0	0.0
Pretest Score	−0.711	0.249	0.073

For both predictive tasks, we evaluated baseline models that only incorporated student gameplay behavior trace data. The results showed that adding facial expressions of emotions data to predictive models of low, medium and high performance groups outperform unimodal models utilizing only gameplay behavior traces or facial expressions of emotions data. However, the unimodal model only utilizing eye-gaze data performed just as well as the multimodal model (ie, gameplay behavior trace and facial expressions). While the unimodal model is more parsimonious and could be chosen as a means to reduce the overall features used, we believe that the multimodal model provides significant value for instructional decision making (eg, scaffolding). Specifically, by incorporating additional modalities, the multimodal model using facial expressions of emotions and gameplay behavior trace has the potential to indicate which aspect of a student's behavior is influencing their performance. For example, if at a given time during a gameplay session, a student is classified to perform poorly on the posttest (ie, "low" performance) based on

Table 5: Interest prediction results (bold values represent best performance)

Data used	*Accuracy*
Majority class baseline	0.344
Gaze	0.541
Face	0.557
Gameplay	0.525
Gaze + Face	0.492
Gameplay + Gaze	**0.590**
Gameplay + Face	0.525
Gameplay + Gaze + Face	0.508

Table 6: Feature coefficients for the gameplay + gaze interest model

Feature	*Low*	*Medium*	*High*
Gameplay-Books	−0.163	−0.012	0.296
Gameplay-Conversation	0.225	−0.250	0.069
Gameplay-Movement	−0.018	−0.143	0.018
Gameplay-PlotPoint	0.0	−0.093	0.107
Gameplay-Posters	−0.025	0.173	−0.014
Gameplay-Scanner	0.175	−0.004	−0.195
Gameplay-Worksheet	0.014	0.021	−0.070
Gameplay-WorksheetSubmit	0.715	0.155	−1.939
GameTime	0.002	0.0	0.0
Gaze-NPC	0.0	0.413	−0.297
Gaze-Food	−0.230	0.010	0.291
Gaze-Lab	0.360	−0.338	0.001
Gaze-Diagnosis	0.697	−0.774	0.001
Gaze-Miscellaneous	−0.057	−0.123	0.132
Gaze-Book	−0.016	0.093	0.009
Gaze-ConceptMatrix	0.344	−0.016	−0.222
Gaze-FixationsPerSecond	−0.015	−0.182	0.263
Pretest Score	0.0	0.176	−0.212

their multimodal data, real-time intervention and scaffolding can be informed by either (or both) of the modalities used in the model.

Instructional decision making that is informed by gameplay behavior traces could provide instructors with opportunity to prompt students to explore unseen parts of the game (eg, focus on relevant materials if the data indicate that the student is focusing on irrelevant features of the game), thus, scaffolding the student to develop a deeper understanding of particular concepts. Scaffolding that is informed by facial expressions of emotions data could be affect-aware, intervening to promote emotional regulation strategies to regulate affective states that could be detrimental to performance such as persistent boredom or frustration. The unimodal model that only incorporates eye-gaze data has the potential to inform instructional decision making about what the student is attending to during game-based learning. The multimodal model utilizing both facial expression and gameplay is preferred to the eye-gaze-only model for real-time scaffolding, because it explains student performance with high accuracy and multiple modalities. We also show that adding eye-gaze data to predictive models of self-reported interest outperforms unimodal models utilizing only gameplay behavior traces or eye-gaze data. These findings highlight

the need for multimodal data and as well as the unique value of different combinations of modalities. The findings also show that utilizing all possible modalities does not always yield the best results, as models using all three modalities are outperformed by combinations of fewer modalities. This is possibly due to larger amounts of noise when using more modalities, affecting the robustness of the predictive model. Redundancy in features across separate modalities can also affect model performance. This finding supports previous multimodal learning analytics work that found occasional redundant and inhibitory effects for certain predictive tasks (D'Mello & Graesser, 2010). An overview of the findings is illustrated in Figure 3, which shows how multimodal data streams extracted from Crystal Island are used as input for predictive modeling of both interest and posttest scores.

For predicting both posttest performance and interest, the multimodal models outperformed the majority classifier by nearly double, with 60.7% and 59.0% accuracy, respectively. We are not aware of previous work that classifies tertiles of either posttest score performance or student interest, making further comparisons difficult. The ultimate goal of this work is to incorporate a predictive model into a real-time system that could adaptively provide feedback to struggling students. Models that incorrectly classify students as low performers or as not interested in the game can negatively influence a student's gameplay experience if the game provides necessary feedback. Alternatively, incorrectly classifying a student as not needing either cognitive or motivational feedback when it is actually needed can also be detrimental. Considering the cost of misclassifications is an important next step of this work and is critical for paving the way toward a real-time scaffolding system.

Predictive models of student posttest scores

To answer RQ1, we investigated how well combinations of student gameplay behavior traces, facial expressions of emotions and eye gaze correctly predict whether students will be low, medium or high performing on the posttest after game-based learning. Our results showed that adding facial expressions of emotions data from the individual action units (AUs) captured during the student's gameplay increased the models' predictive accuracy. This combination of modalities outperformed the baseline model, with models incorporating all modalities achieving lower

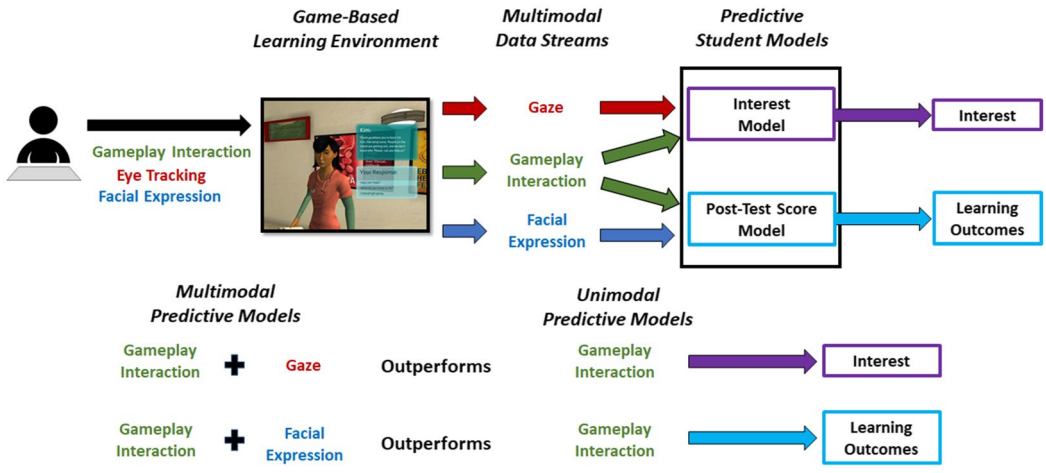

Figure 3: Overview of the multimodal data streams and predictive modeling approach

performance in comparison, potentially due to feature redundancy and noise. Additionally, the model that utilized only student gaze data performed equally as well as the Gameplay + Face model. A potential explanation for this result is the redundancy of eye tracking features and gameplay features, causing the logistic regression model using Gameplay + Gaze to perform poorly. The gaze features are categorized by the names of the in-game objects, while the gameplay features are often categorized by the actions students take in the game. These actions are often parameterized by the individual objects within the game, such as reading a specific virtual book or scanning a specific food item. Thus, there is overlap when using these features to predict posttest performance. Engineering new fine-grained feature representations that do not completely overlap for both the gaze features and the gameplay features could provide new insight into the prediction of student posttest scores. However, the combination of gameplay and facial expression for predicting student posttest performance is consistent with previous research (Sawyer *et al.*, 2017). Gameplay data account for basic behavior in the game-based learning environment and facial expression data captures emotion-related responses of students to their interactions within the environment. Prior work has shown that emotion is related to learning in many ways and this work reinforces those findings (Grafsgaard, Wiggins, Boyer, Wiebe, & Lester, 2014; Pardos, Baker, San Pedro, Gowda, & Gowda, 2014). However, we do not draw any conclusions for specific emotions and their relationship to learning. Recent work has questioned the validity of using facial movement to label emotions, even for people within a single situation, such as students interacting with the same game-based learning environment (Barrett, Adolphs, Marsella, Martinez, & Pollak, 2019). The emotion labels assigned by automated systems often do not consider the fact that people express emotions differently based on aspects such as socio-cultural and environmental factors. To avoid this issue, we use the individual facial movements (AUs) directly as features, which compose the various affective labels in FACET. We can then observe the variability of facial movements of specific students. The strongest predictors for posttest scores in the Gameplay + Face model were Gameplay-Posters, Gameplay-Conversation, Face-AU15, Face-AU28, Face-AU14, Face-AU4, Face-AU20, Face-AU43 and Pretest Score. The Gameplay-Conversation feature indicates a negative relationship with posttest score, especially for students in the medium posttest performance category. Notably, *AU15 Lip Corner Depressor*, *AU14 Dimpler* and *AU43 Eyes Closed* have been shown to be predictive of learning in previous work (Grafsgaard, Wiggins, Vail, *et al.*, 2014; Sawyer *et al.*, 2017). *AU14 Dimpler* specifically showed a negative relationship with learning, which is in line with this research.

Predictive models of student interest

To answer RQ2, we investigated how well combinations of student gameplay, facial expressions of emotions and eye-gaze data classified low, medium and high interest groups of students after game-based learning. Our results showed that the most accurate model utilized gameplay and gaze features, which yielded higher predictive accuracy compared to models that used only a single modality. Additionally, adding facial expression features to this model decreased performance. Student gameplay data have value for predicting student interest because it provides a window into students' problem solving in the game and whether they are interacting with the game's artifacts. Students who engage in behaviors that are off-task (eg, solitary behavior, inactivity, gaming the system; see Sabourin, Rowe, Mott, & Lester, 2013) may be disengaged, which can be captured from gameplay. Student gaze data can also indicate engagement. For example, a student who is frequently scanning a food item while primarily looking at other unrelated objects in the game could be identified as disengaged or potentially unmotivated. This situation could indicate the student is either struggling to grasp the material or perhaps they are "gaming the system" through an exhaustive approach (Sabourin *et al.*, 2013). We found that by adding in facial expression

data, the performance of the predictive models of interest decreased. This could be due to the lack of context of the facial expression data. For instance, knowing when students expressed certain facial expressions (eg, during reading) as opposed to a summary could be meaningful for predicting interest. The most predictive features in our models of student interest were the following: Gameplay Books, Gameplay-Conversation, Gameplay-Movement, Gameplay-Posters, Gameplay-Scanner, Gameplay-WorksheetSubmit, Gaze-NPC, Gaze-Food, Gaze-Lab, Gaze-Diagnosis, Gaze-Miscellaneous, Gaze-ConceptMatrix and Pretest Score. Notably, the Gameplay-Conversation, Gameplay-Scanner and Gameplay-WorksheetSubmit features were previously found to be predictive of interest (Sawyer *et al.*, 2018). The number of conversations has a negative relationship with interest, possibly because it could indicate that the student is not grasping the material and is engaging with off-task behavior. The number of scanner events has a negative relationship with interest, which could be due to students who are repeatedly scanning objects to finish the mystery without expending effort. However, more thoughtful scanning and hypothesis testing could also indicate higher interest. Similarly, students who look more frequently at their diagnosis worksheet and submit the worksheet more frequently tend to have lower interest, again possibly due to a lack of thoughtful testing.

Limitations

While this study showed the promise of using multimodal data to model student posttest score performance and self-reported interest in game-based learning, there are several limitations of this work. First, in measuring and coding students' posttest performance and interest, we grouped the students into groups of low, medium and high based on their scores. The splits are atheoretical and purely data-determined, as are any percentile-based grouping approaches, but they do provide insight into relative differences between students. Second, the instruments we used to measure student knowledge and interest can only measure these characteristics at one point in time. In reality, student knowledge and interest change over time during gameplay. It is important to note that the IMI is retrospective in nature, so it measures students' intrinsic motivation for playing the game following their interaction with it. An alternative approach could have been to prompt the student for this measure throughout gameplay; however, interrupting the student's gameplay to repeatedly administer the IMI would potentially have been disruptive to learning and decrease engagement and the overarching goal of this work is to create predictive models that operate unobtrusively. Third, the feature representation of gameplay behavior traces, facial expressions of emotions and eye gaze are nontemporal; rather than explicitly encoding time as a variable, they implicitly encode time and explicitly encode an aggregate summary of data up to a particular moment in time. This representation loses information, as patterns of these data streams likely emerge over time. Using the sequence of these data streams would be an alternative approach, but this would require a machine learning model that accommodates sequential data. Finally, our study only focused on three modalities of student data. Utilizing additional channels of data could further improve performance.

Conclusion and future work

Advances in sensor-based technologies introduce the opportunity to leverage both facial expression and gaze data streams in game-based learning environments. Together with gameplay behavior traces, these can enable predictive models to achieve high predictive accuracy for both student posttest performance and interest. Predictive models that have access to this information could thereby guide real-time scaffolding in game-based learning. Prior work has shown that both facial expression and gaze have shown promise for predicting learning outcomes, but little work has investigated the potential synergistic effects of employing multiple modalities for

these prediction tasks. To investigate the potential of multimodal predictive modeling, we created predictive models equipped with combinations of gameplay, gaze and facial expression data to predict student posttest performance and interest in a game-based learning environment. We created different predictive models using each possible combination of available modalities and evaluated their respective predictive accuracies. The results demonstrate that models utilizing gameplay and facial expression data outperform models that only utilize gameplay data when predicting student posttest performance. Additionally, models using a combination of gameplay and facial expression outperform models using only facial expression data. For predicting interest, models utilizing gameplay and gaze data outperform models using only gameplay data. Likewise, models utilizing both gameplay and gaze data outperform models using only eye gaze data.

We found that when predicting learning, gaze-only models performed just as well as models incorporating both gameplay and facial expression. However, when adding gaze data to the gameplay and facial expression model, performance decreases. This is possibly due to the redundancy of the feature representation of the gaze data. Many of the features in the gameplay modality are parameterized by the individual in-game objects, which are equivalent to the gaze-based features. Predicting student posttest scores using facial expression and gameplay is advantageous over only using gaze because it offers deeper insight about a student's performance. Scaffolding informed by more than one modality has significant promise for providing real-time, individualized feedback that addresses the specific issues a student may be experiencing.

The findings suggest several promising directions for future work. First, it will be important to investigate its generalizability by studying other learning environments using different combinations of modalities and with different student populations. Second, it will be important to explore more expressive feature representations for each modality. We used a static-based representation of each modality in the work reported here to examine the relationships between each modality with both student posttest performance and interest after game-based learning. A promising alternative could be a sequential representation that adds even more granularity to the predictive modeling approach. This sequential model could be updated with a more accurate estimation of student posttest performance and interest and it could be used to evaluate early predictions of both student posttest performance and interest. These early predictions could then be used to adapt the game at an early point. Evaluating how predictions improve over time as the models incorporate more student data will help determine when predictions are reliable. Third, it will be instructive to investigate more flexible modeling techniques to achieve higher performance. With the insight of which features perform well in both the predictive tasks we presented, more sophisticated modeling techniques may be able to achieve even higher accuracy. Fourth, utilizing additional modalities, such as posture and galvanic skin response, may further increase predictive accuracy. Fifth, we did not measure interest during gameplay. An alternative could be to measure interest at different time points in student gameplay and then, model relationships between multimodal interactions and current interest. A final promising direction is to investigate how multimodal models can be incorporated into game-based learning environments to support advanced learning analytics for real-time adaptive scaffolding.

Acknowledgements

This study was supported by the funding from the Social Sciences and Humanities Research Council of Canada (SSHRC 895-2011-1006). 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.

Statements on Open Data, Ethics and Conflict of Interest

Due to human subject protection policies, the study data are not open.

This study was conducted with the IRB approval of North Carolina State University.

There is no potential conflict of interest in this work.

References

Ainley, M., & Ainley, J. (2019). Motivation and learning: Measures and methods. In K. A. Renninger & S. E. Hidi (Eds.), *The Cambridge Handbook of Motivation and Learning*. Cambridge, UK: Cambridge University Press.

Akram, B., Min, W., Wiebe, E., Mott, B., Boyer, K. E., & Lester, J. (2018). Improving stealth assessment in game-based learning with LSTM-based analytics. In K. E. Boyer & M. Yudelson (Eds.), *Proceedings from EDM'18: Eleventh International Conference on Educational Data Mining* (pp. 208–218). Boston, MA: International Educational Data Mining.

Alonso-Fernández, C., Cano, A. R., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Lessons learned applying learning analytics to assess serious games. *Computers in Human Behavior*, *99*, 301–309.

Alonso-Fernández, C., Martínez-Ortiz, I., Caballero, R., Freire, M., & Fernández-Manjón, B. (2020). Predicting students' knowledge after playing a serious game based on learning analytics data: A case study. *Journal of Computer Assisted Learning*, *36*(3), 350–358.

Aslan, S., Alyuz, N., Tanriover, C., Mete, S. E., Okur, E., D'Mello, S. K., & Arslan Esme, A. (2019). Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms. In *Proceedings from CHI'19: Conference on Human Factors in Computing Systems* (pp. 1–12). New York, NY: Association for Computing Machinery.

Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: Issues and challenges. *Computers in Human Behavior*, *96*, 207–210.

Azevedo, R., Taub, M., & Mudrick, N. V. (2018). Using multi-channel trace data to infer and foster self-regulated learning between humans and advanced learning technologies. In D. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 254–270). New York, NY: Routledge.

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*, 223–241.

Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*, 423–443.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*, 1–68.

Blikstein, P. (2013). Multimodal learning analytics. In D. Suthers, K. Verbert, E. Duval, & X. Ochoa (Eds.), *Proceedings from LAK'13: Third International Conference on Learning Analytics and Knowledge* (pp. 102–106). New York, NY: Association for Computing Machinery.

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, *3*, 220–238.

Botelho, A. F., Baker, R. S., & Heffernan, N. T. (2017). Improving sensor-free affect detection using deep learning. In E. André, R. Baker, X. Hu, M. Rodrigo, & B. du Boulay (Eds.), *Proceedings from AIED'17: International Conference on Artificial Intelligence in Education* (pp. 40–51). Cham, Switzerland: Springer International Publishing.

Cloude, E. B., Taub, M., Lester, J., & Azevedo, R. (2019). The role of achievement goal orientation on metacognitive process use in game-based learning. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Proceeding from AIED'19: International Conference on Artificial Intelligence in Education* (pp. 36–40). Cham, Switzerland: Springer International Publishing.

D'Mello, S. K., & Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, *20*, 147–187.

Dever, D. A., & Azevedo, R. (2019). Autonomy and types of informational text presentations in game-based learning environments. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Proceedings from AIED'19: International Conference on Artificial Intelligence in Education* (pp. 110–120). Cham, Switzerland: Springer International Publishing.

Dever, D. A., Wiedbusch, M., & Azevedo, R. (2019). Learners' gaze behaviors and metacognitive judgments with an agent-based multimedia environment. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Proceedings from AIED'19: International Conference on Artificial Intelligence in Education* (pp. 58–61). Cham, Switzerland: Springer International Publishing.

Di Mitri, D., Schneider, J., Specht, M., & Drachsler, H. (2018). From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning*, *34*, 338–349.

D'Mello, S., Dieterle, E., & Duckworth, A. (2017). Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational Psychologist*, *52*, 104–123.

D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies*, *70*, 377–398.

du Boulay, B., & Del Soldato, T. (2016). Implementation of motivational tactics in tutoring systems: 20 years on. *International Journal of Artificial Intelligence in Education*, *26*, 170–182.

Ekman, P., & Rosenberg, E. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)*. New York, NY: Oxford University Press.

Emerson, A., Sawyer, R., Azevedo, R., & Lester, J. (2018). Gaze-enhanced student modeling for game-based learning. In *Proceedings from UMAP'18: Twenty-Sixth International Conference on User Modeling, Adaptation, and Personalization* (pp. 63–72). New York, NY: Association for Computing Machinery.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*, 59–109.

Geden, M., Emerson, A., Rowe, J., Azevedo, R., & Lester, J. (2020). Predictive student modeling in educational games with multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(1), 654–661.

Giannakos, M. N., Sharma, K., Pappas, I. O., Kostakos, V., & Velloso, E. (2019). Multimodal data as a means to understand the learning experience. *International Journal of Information Management*, *48*, 108–119.

Gomes, J., Yassine, M., Worsley, M., & Blikstein, P. (2013). Analysing engineering expertise of high school students using eye tracking and multimodal learning analytics. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings from EDM'13: International Conference on Educational Data Mining* (pp. 375–377). Boston, MA: International Educational Data Mining.

Grafsgaard, J., Wiggins, J., Boyer, K. E., Wiebe, E., & Lester, J. (2014). Predicting learning and affect from multimodal data streams in task-oriented tutorial dialogue. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings from EDM'14: International Conference on Educational Data Mining* (pp. 122–129). Boston, MA: International Educational Data Mining.

Grafsgaard, J. F., Wiggins, J. B., Vail, A. K., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2014). The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In *Proceedings of the Sixteenth International Conference on Multimodal Interaction* (pp. 42–49). New York, NY: Association for Computing Machinery.

Henderson, N., Emerson, A., Rowe, J., & Lester, J. (2019). Improving sensor-based affect detection with multimodal data imputation. In *Proceedings from ACII'19: Eighth International Conference on Affective Computing and Intelligent Interaction* (pp. 669–675). New York, NY: IEEE.

Henderson, N. L., Rowe, J. P., Mott, B. W., Brawner, K., Baker, R., & Lester, J. C. (2019). 4D affect detection: Improving frustration detection in game-based learning with posture-based temporal data fusion. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Proceedings from AIED'19: International Conference on Artificial Intelligence in Education* (pp. 144–156). Cham, Switzerland: International Springer Publishing.

Hidi, S. E., & Renninger, K. A. (2019). Interest development and its relation to curiosity: Needed neuroscientific research. *Educational Psychology Review*, *31*, 1–20.

Hutt, S., Krasich, K., Mills, C., Bosch, N., White, S., Brockmole, J. R., & D'Mello, S. K. (2019). Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction*, *29*, 821–867.

iMotions. (2016). *Attention Tool, 6.2*. Boston, MA: iMotions Inc.

Lallé, S., Conati, C., & Azevedo, R. (2018). Prediction of student achievement goals and emotion valence during interaction with pedagogical agents. In *Proceedings from AAMAS'18: Seventeenth International Conference on Autonomous Agents and MultiAgent Systems* (pp. 1222–1231). New York, NY: Association for Computing Machinery.

Lane, H. C., & D'Mello, S. K. (2019). Uses of physiological monitoring in intelligent learning environments: A review of research, evidence, and technologies. In T. Parsons, L. Lin, & D. Cockerham (Eds.), *Mind, Brain and Technology. Educational Communications and Technology: Issues and Innovations* (pp. 67–86). Cham, Switzerland: Springer.

Lee, J. Y., Donkers, J., Jarodzka, H., & van Merriënboer, J. J. (2019). How prior knowledge affects problem-solving performance in a medical simulation game: Using game-logs and eye-tracking. *Computers in Human Behavior*, *99*, 268–277.

Loderer, K., Pekrun, R., & Lester, J. C. (2018). Beyond cold technology: A systematic review and meta-analysis on emotions in technology-based learning environments. *Learning and Instruction*, 101162.

Mayer, R. E. (2019). Computer games in education. *Annual Review of Psychology*, *70*, 531–549.

McAuley, E., Duncan, T., & Tammen, V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, *60*, 48–58.

Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Smith, A., Wiebe, E., … Lester, J. C. (2020). (in press). DeepStealth: Game-based learning stealth assessment with deep neural networks. *IEEE Transactions on Learning Technologies*, *13*(2), 312–325.

Min, W., Mott, B., Rowe, J., Taylor, R., Wiebe, E., Boyer, K. E., & Lester, J. (2017). Multimodal goal recognition in open-world digital games. In *Proceedings from AIIDE'17: Thirteenth Annual AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (pp. 80–86).

Nietfeld, J. L., Shores, L. R., & Hoffmann, K. F. (2014). Learning environment self-regulation and gender within a game-based learning environment. *Journal of Educational Psychology*, *106*, 961–973.

Ochoa, X., & Worsley, M. (2016). Augmenting learning analytics with multimodal sensory data. *Journal of Learning Analytics*, *3*, 213–219.

Oviatt, S., Grafsgaard, J., Chen, L., & Ochoa, X. (2018). Multimodal learning analytics: Assessing learners' mental state during the process of learning. In S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, & A. Kruger (Eds.), *The Handbook of Multimodal-Multisensor Interfaces* (pp. 331–374). Association for Computing Machinery and Morgan & Claypool.

Pardos, Z. A., Baker, R. S., San Pedro, M. O., Gowda, S. M., & Gowda, S. M. (2014). Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics*, *1*, 107–128.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, *50*, 258–283.

Plass, J. L., Mayer, R. E., & Homer, B. D. (2020). *Handbook of game-based learning*. Cambridge, MA: MIT Press.

Qian, M., & Clark, K. R. (2016). Game-based Learning and 21st century skills: A review of recent research. *Computers in Human Behavior*, *63*, 50–58.

Rajendran, R., Kumar, A., Carter, K. E., Levin, D. T., & Biswas, G. (2018). Predicting learning by analyzing eye-gaze data of reading behavior. In K. E. Boyer & M. Yudelson (Eds.), *Proceedings of the International Conference on Educational Data Mining. Proceedings from EDM'18: Eleventh International Conference on Educational Data Mining* (pp. 455–461). Boston, MA: International Educational Data Mining.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372–422.

Rotgans, J. I., & Schmidt, H. G. (2011). Situational interest and academic achievement in the active-learning classroom. *Learning and Instruction*, *21*, 58–67.

Rowe, J., Shores, L., Mott, B., & Lester, J. (2011). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education*, *21*, 115–133.

Ryan, R. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, *43*, 450–461.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68–78.

Sabourin, J. L., Rowe, J. P., Mott, B. W., & Lester, J. C. (2013). Considering alternate futures to classify off-task behavior as emotion self-regulation: A supervised learning approach. *Journal of Educational Data Mining*, 5, 9–38.

Sawyer, R., Rowe, J., Azevedo, R., & Lester, J. (2018). Modeling player engagement with Bayesian hierarchical models. In *Proceedings from AIIDE'18: Fourteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (pp. 215–221).

Sawyer, R., Smith, A., Rowe, J., Azevedo, R., & Lester, J. (2017). Enhancing student models in game-based learning with facial expression recognition. In *Proceedings from UMAP'17: Twenty-Fifth Conference on User Modeling, Adaptation and Personalization* (pp. 192–201). New York, NY: Association for Computing Machinery.

Sharma, K., Papamitsiou, Z., & Giannakos, M. (2019). Building pipelines for educational data using AI and multimodal analytics: A "grey-box" approach. *British Journal of Educational Technology*, *50*, 3004–3031.

Sharma, K., Papamitsiou, Z., Olsen, J. K., & Giannakos, M. (2020, March). Predicting learners' effortful behaviour in adaptive assessment using multimodal data. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 480–489).

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, *63*, 106–117.

Spires, H. A., Rowe, J. P., Mott, B. W., & Lester, J. C. (2011). Problem solving and game-based learning: Effects of middle grade students' hypothesis testing strategies on learning outcomes. *Journal of Educational Computing Research*, *44*, 453–472.

Taub, M., Azevedo, R., Rajendran, R., Cloude, E. B., Biswas, G., & Price, M. J. (2019). (in press). How are students' emotions related to the accuracy of cognitive and metacognitive processes during learning with an intelligent tutoring system? *Learning and Instruction*, 101200.

Taub, M., Mudrick, N. V., Azevedo, R., Millar, G. C., Rowe, J., & Lester, J. (2017). Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with Crystal Island. *Computers in Human Behavior*, *76*, 641–655.

Taub, M., Sawyer, R., Lester, J., & Azevedo, R. (2020). The impact of contextualized emotions on self-regulated learning and scientific reasoning during learning with a game-based learning environment. *International Journal of Artificial Intelligence in Education*, *30*, 97–120.

Taub, M., Sawyer, R., Smith, A., Rowe, J., Azevedo, R., & Lester, J. (2020). The agency effect: The impact of student agency on learning, emotions, and problem-solving behaviors in a game-based learning environment. *Computers & Education*, *147*, 1–19.

Tsai, M. J., Huang, L. J., Hou, H. T., Hsu, C. Y., & Chiou, G. L. (2016). Visual behavior, flow and achievement in game-based learning. *Computers & Education*, *98*, 115–129.

Vlachopoulos, D., & Makri, A. (2017). The effect of games and simulations on higher education: A systematic literature review. *International Journal of Educational Technology in Higher Education*, *14*, 1–33.

Vrzakova, H., Amon, M. J., Stewart, A., Duran, N. D., & D'Mello, S. K. (2020, March). Focused or stuck together: Multimodal patterns reveal triads' performance in collaborative problem solving. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 295–304).

Wentzel, K. R., & Miele, D. B. (2016). *Handbook of motivation at school* (2nd ed.). Abingdon, UK: Routledge.