JOURNAL OF LEARNING ANALYTICS

Volume 9(3), 75–87. https://doi.org/10.18608/jla.2022.7631

Multimodal Data Fusion to Track Students' Distress during Educational Gameplay

Jewoong Moon[1], Fengfeng Ke[2], Zlatko Sokolikj[3], and Ibrahim Dahlstrom-Hakki[4]

Abstract

Using multimodal data fusion techniques, we built and tested prediction models to track middle-school student distress states during educational gameplay. We collected and analyzed 1,145 data instances, sampled from a total of 31 middle-school students' audio- and video-recorded gameplay sessions. We conducted data wrangling with student gameplay data from multiple data sources, such as individual facial expression recordings and gameplay logs. Using supervised machine learning, we built and tested candidate classifiers that yielded an estimated probability of distress states. We then conducted confidence-based data fusion that averaged the estimated probability scores from the unimodal classifiers with a single data source. The results of this study suggest that the classifier with multimodal data fusion improves the performance of tracking distress states during educational gameplay, compared to the performance of unimodal classifiers. The study finding suggests the feasibility of multimodal data fusion in developing game-based learning analytics. Also, this study proposes the benefits of optimizing several methodological means for multimodal data fusion in educational game research.

Notes for Practice (research paper)

- Distress is a cognitive-affective state related to task-specific frustration that involves unpleasant feelings.
- Tracking students' cognitive-affective states during educational gameplay drives timely and adaptive learning support.
- Multimodal learning analytics gathers learning-related insights from students' multimodal interactions.
- Using multimodal data fusion can improve the prediction performance in tracking cognitive-affective states when building game-based multimodal learning analytics.

Keywords

Multimodal data fusion, educational game, distress, educational data mining

Submitted: 24/11/21 — **Accepted:** 21/06/22 — **Published:** 21/11/22

Corresponding author [1] Email: jmoon19@ua.edu Address: Department of Educational Leadership, Policy, and Technology Studies, The University of Alabama, Tuscaloosa, United States. ORCID ID: https://orcid.org/0000-0001-6311-3019

[2] Email: fke@fsu.edu Address: Department of Educational Psychology and Learning Systems, Florida State University, Tallahassee, United States. ORCID ID: https://orcid.org/0000-0003-4203-1203

[3] Email: zs09f@fsu.edu Address: Department of Scientific Computing, Florida State University, Tallahassee, United States. ORCID ID: https://orcid.org/0000-0002-3090-4460

[4] Email: ibrahim_dahlstrom-hakki@terc.edu Address: TERC, Cambridge, United States. ORCID ID: https://orcid.org/0000-0002-9418-1900

1. Introduction

Educational gameplay typically promotes motivation and immersion (Filsecker & Hickey, 2014). During educational gameplay, students experience dynamic cognitive-affective states. Students engaging in game tasks often experience the state of distress, one of the frequent cognitive-affective states that occur during gameplay. Distress refers to unpleasant feelings or emotions when individuals are cognitively overwhelmed. In educational game research, distress connotes student task-related frustration (Yeh et al., 2015). For example, student wheel-spinning experiences during gameplay can give rise to a high level of frustration, whereas their successful gameplay performance can result in high levels of engagement. During gameplay, the dynamics of distress change or mediate learner perseverance (Silvervarg et al., 2018) and problem-solving performance (Eseryel et al., 2014). Hence, researchers have sought to track students' various cognitive-affective states to enable adaptive support during educational gameplay (Plass et al., 2019; Rodrigo, 2011).

Despite emerging studies of cognitive-affective states in educational gameplay, previous research has limitations. First, existing research shows limited performance in tracking the state of distress during educational gameplay. Distress is particularly indicative of student difficulty or task challenges in gameplay (Bowman & Tamborini, 2012). Hence, tracking evidence of student distress in gameplay can drive adaptive learning supports. Although distress is a good indicator of negative cognitive-affective states during gameplay, prior research rarely explores how to computationally track it. Second, previous studies highlight the limited success of singular measures of tracking students' cognitive-affective states. However, sparse research addresses how to systematically collect, process, and analyze multimodal data when tracking cognitive-affective states during educational gameplay (Ochoa & Worsley, 2016). Few empirical studies explore ways to synthesize multimodal data during student gameplay, warranting a search for ways to effectively manage multimodal data and to track cognitive-affective states. Accordingly, we propose multimodal data fusion using various computational measures. This study focuses on the design and conduct of multimodal data fusion when tracking student states of distress during educational gameplay.

1.1. Distress in Educational Gameplay

In educational game research, tracking the dynamics of students' cognitive-affective states is important (Baker et al., 2010; Rodrigo, 2011). During gameplay, students experience dynamic cognitive-affective states (Shute et al., 2015) that are strongly related to their performance (LePine et al., 2004), self-regulation, and problem-solving attitudes (Taub et al., 2020). Researchers identified various cognitive-affective states associated with gameplay performance and academic outcomes during educational gameplay, reporting that such cognitive-affective states are predictive of student learning outcomes in educational gameplay. Hamari et al. (2016) discovered how various cognitive-affective states (i.e., flow, engagement, and immersion) are associated with learning achievement. Shute et al. (2015) also investigated how knowledge comprehension, persistence, affective states, and in-game performance influence student learning during educational gameplay. Their study findings suggest that cognitive-affective states strongly relate to gameplay performance and learning achievement.

Students largely undergo distress as one of the cognitive-affective states in educational gameplay. In the literature, stress generally refers to a cognitive-affective state that involves emotional and psychological changes with bodily responses (Chrousos et al., 2013). Distress, as a type of stress, generally occurs when individuals experience exhaustion that lowers their energy (LePine et al., 2004). Specifically, distress is an unpleasant state associated with significant negative consequences as a result of task failures (Fultz et al., 1988; Naylor, 2020). Distress features a mixture of symptoms involving depression, anxiety, and anger (Kessler et al., 2002). Rudland et al. (2020) state that distress explains individuals' negative affective states as a result of stress. They argue that distress relates to students' perceived task difficulty and learning performance. Hence their work indicates that distress is a precursor of student learning outcomes.

In the context of educational gameplay, distress can be associated with wheel-spinning processes, involving unproductive gameplay consequences and overwhelming gaming tasks (Beck & Rodrigo, 2014). Wheel-spinning refers to a behavioural state of exerting unproductive effort in which students use too much time in struggling with game tasks. Once a student undergoes repeated failures, they are likely to be overwhelmed and become distressed. There are two types of cognitive-affective states that students experiencing wheel-spinning can face: frustration and distress. Whereas frustration involves a high degree of arousal and negative valence in gameplay, distress largely relates to demotivation and excessive levels of anxiety. Frustration refers to a student's mental exertion to overcome impasses (Henderson et al., 2019), whereas distress is largely associated with repeated and unproductive failure cycles that result in quitting behaviours during educational gameplay. Therefore, tracking students' distress states is useful in understanding when and how to adaptively deliver in-game learning support. Despite existing studies that discuss a variety of cognitive-affective states in educational game research, few seek ways to track students' states of distress.

1.2. Detection of Cognitive-Affective States in Educational Game Research

Because cognitive-affective states indicate students' perceived difficulty and academic achievement, detecting cognitive-affective states is essential for driving adaptive learning support in digital learning environments. Researchers suggest the potential of using data analytics to track dynamic changes in cognitive-affective states (Baker et al., 2010; Verma et al., 2020). Rodrigo (2011) — for example, quantitative field observations and transition likelihood metrics — to examine transition paths of different cognitive-affective states. In addition, recent research introduces the design and development of automatic detectors for tracking student cognitive-affective states in digital learning environments (Bosch et al., 2014; Chen et al., 2020). For instance, DeFalco et al. (2018) explore ways to detect student frustration in a serious game. They use machine-learning algorithms and behaviour observation records to develop affect detectors of frustration and engagement. Bosch et al. (2014) built an automatic detector by using individuals' facial features with machine learning, developing classification models to track various cognitive-affective states (e.g., boredom, flow, engagement). As another example, Munshi et al. (2018) used affect data, learning outcomes, and action log data to analyze students' cognitive-affective states in the intelligent tutoring

system Betty's Brain. The result of this study shows how certain affective states (i.e., frustration and boredom) are associated with either positive or negative learning performance. Bosch et al. (2015) built a supervised machine learning model by using student facial expression data. Using FACET, a software to track facial data, they analyzed students' action units (AUs) and then built a prediction model tracking major affective states.

Despite advances in automatic-detector design in digital learning environment research, existing studies have limitations. Educational games are highly interactive digital learning platforms that require learners to use various modalities. Therefore, collecting student unimodal data during educational gameplay may not comprehensively detect the dynamics of their cognitive-affective states. Given that humans generally express cognitive-affective states by a combination of multiple sensory cues (Ochoa & Worsley, 2016), researchers should collect and synthesize data from multiple sources to acquire a comprehensive understanding of cognitive-affective state changes. Responding to this need, researchers suggest the importance of multimodal learning analytics (MmLA), a data analytics discipline that focuses on systematic data collection, analysis, and representation of multimodal data to better denote student learning. MmLA also considers data granularity to precisely detect the dynamics of cognitive-affective states in relation to student learning. The promise of MmLA increases researcher attention to integrating multimodal data mining into educational game research.

A recent study by Emerson et al. (2020) conducts multimodal predictive modelling by integrating three different types of data (gameplay behaviour traces, facial expressions of emotions, and eye-gaze data). Their multimodal fused model was superior at predicting student interests and knowledge about microbiology. Henderson et al. (2020) explore multiple multimodal data fusion approaches to detect affective states during gameplay. They use and compare feature-level and decision-level data fusions to yield better prediction performance of multiple affective states (i.e., bored, confused, concentrated, frustrated, and surprised).

There are still knowledge gaps that should be resolved. First, despite the majority of studies tracking various cognitive-affective states in digital learning environments, limited research has explored ways to detect students' distress states in a game-based learning context. Even though distress is a potential precursor of in-game learning performance, existing approaches rarely developed and tested models to track distress states. Second, a distress-tracking model using multimodal data fusion has not been explored. Although existing literature echoes the effectiveness of multimodal data fusion in digital learning environments, its usefulness in tracking distress states during game-based learning has not been investigated. The purpose of this study, therefore, is to design, implement, and evaluate multimodal data fusion measuring students' state of distress during educational gameplay. The research question of this study is this: To what extent does multimodal data fusion improve the detection of students' distress state during educational gameplay?

2. Methods

2.1. Data

We collected data from thirty-one middle-school students (Grades 6–8) engaged in 1–1.5-hour gameplay sessions. Study participants played Zoombinis either in school classrooms or in their homes. During their play, study participants' postures, including facial expressions, and gameplay screens were recorded. Students played Zoombinis (Asbell-Clarke et al., 2021), a 2D puzzle game aimed at promoting young children's computational thinking development (Figure 1). For data analysis, we sampled and analyzed student data from the game puzzle Mudball Wall in Zoombinis. This puzzle requires students to use strategies to find accurate patterns of shapes and colours of mudballs to hit highlighted cells on the wall and then toss Zoombinis characters onward. We collected multiple types of data: 1) video recordings of student facial expressions and gameplay screens during Zoombinis play, and 2) gameplay logs that captured gaming behaviours, strategies, and performance results. We used RapidMiner data-mining software (Hofmann & Klinkenberg, 2016) to process and analyze the data.

2.2. Data Wrangling

With the collected multimodal data (i.e., facial-expression data and gameplay logs), we implemented multimodal data fusion — a data-mining technique blending different types or modalities of data — to predict changes in the outcome variable, which was the state of distress. The state of distress was coded as a dichotomous result with 0 or 1. For multimodal data fusion, we implemented a decision-level fusion approach. Compared to feature-level fusion, which directly merges feature vectors, decision-level fusion integrates the predicted scores of each classifier (Alam et al., 2015) by categorizing results for each modality before fusing the outcome results. For the decision-level fusion, we combined the average confidence values of sub-decisions into the final decision about the target outcome measure.

This study's multimodal data fusion contains four data-wrangling processes. Figure 2 depicts our application of multimodal data fusion. First, we gathered multi-channel data from students, including their facial data from two facial-expression detection toolkits (OpenFace and Facial Expression Recognition [FER-2013]) as well as Zoombinis gameplay logs. OpenFace generates

AU data (Baltrušaitis et al., 2016), tracking facial-muscle movements, whilst a FER-2013 data-driven open-source toolkit computes the probabilities of the "big five" emotions, based on image-based emotion classification data (FER-2013; Goodfellow et al., 2013). FER-2013 data[1] comprises a total of 35,887 48 × 48 pixel grayscale facial images. This dataset was built by Pierre Luc Carrier and Aaron Courville and introduced at the ICML 2013 workshop's facial expression recognition competition. During the data fusion, we also included student gameplay logs with their total gameplay performance statistics and in-game actions. We merged the metadata of these three different data sources and rearranged it by gameplay rounds and trials. We finalized and used for model development a total of 1,145 data instances.

Figure 1. Gameplay scene (Mudball puzzle)

This study's multimodal data fusion contains four data-wrangling processes. Figure 2 depicts our application of multimodal data fusion. First, we gathered multi-channel data from students, including their facial data from two facial-expression detection toolkits (OpenFace and Facial Expression Recognition [FER-2013]) as well as Zoombinis gameplay logs. OpenFace generates AU data (Baltrušaitis et al., 2016), tracking facial-muscle movements, whilst a FER-2013 data-driven open-source toolkit computes the probabilities of the "big five" emotions, based on image-based emotion classification data (FER-2013; Goodfellow et al., 2013). FER-2013 data[2] comprises a total of 35,887 48 × 48 pixel grayscale facial images. This dataset was built by Pierre Luc Carrier and Aaron Courville and introduced at the ICML 2013 workshop's facial expression recognition competition. During the data fusion, we also included student gameplay logs with their total gameplay performance statistics and in-game actions. We merged the metadata of these three different data sources and rearranged it by gameplay rounds and trials. We finalized and used for model development a total of 1,145 data instances.

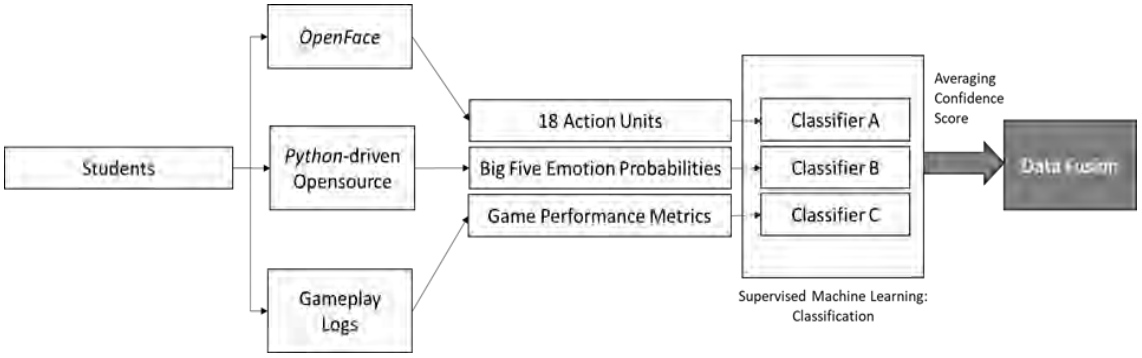

Figure 2. Multimodal data fusion

[1] https://www.kaggle.com/datasets/msambare/fer2013
[2] https://www.kaggle.com/datasets/msambare/fer2013

Second, we observed and coded recorded gameplay video footage to perform ground-truth labelling of student distress. Two professional coders independently labelled the presence or state of distress across the entire dataset and then reconciled their results until they approached 100% interrater agreement. The two coders were equally familiar with Zoombinis gameplay at the same level. Using 20% of the entire set of gameplay recordings, we developed a coding scheme, outlined in Table 1, to operationalize the occurrence of distress during student gameplay. Specifically, two coders watched a collection of example gameplay videos and brainstormed the behaviour examples indicating the moments of distress. The highest inter-reliability among the two coders was .81 on average. The coders then iteratively checked for inconsistent coding results and resolved them by discussion. We regularly conducted data sessions for coder training until reaching the inter-reliability level above .80.

Using the refined coding scheme, we conducted the ground-truth labelling by the coders and matched coding results until they reached 100% mutual agreement. We built our coding scheme based on prior research on behavioural examples of distress states. Previous research reports that individuals are likely to show a group of non-verbal behaviours relating to their distress states. For instance, Biglan (1991) stated that distressed behaviours generally contain a list of non-verbal behaviour cues, such as aggressive and sombre facial expressions, tense eyelids, lips pressed tightly together, and slamming something. Using sample videos of student gameplay, we tested the initial coding scheme and then iteratively refined distress-related behaviour examples to be coded. We finally merged ground-truth labelling results with all gameplay data extracted from the three data channels.

Third, we separately executed machine-learning classifiers with the three data sources (i.e., AUs from OpenFace, estimated emotion probabilities from FER data-driven open source, and gameplay logs) based on the expert-labelling results. Here we computed each machine-learning classifier's probability scores for the state of distress. We then used confidence-based fusion (Alam et al., 2015) to create a classifier. We computed sub-decisions by confidence scores, using independent classifiers with different data sources. We then computed the overall prediction results of the state of distress, by averaging the confidence ratings from three independent data sources with a pre-set threshold condition (i.e., higher than 50%).

Last, we compared the performance of the classifiers with individual data sources and the one with the fused data, to see if and how multimodal data fusion enhanced the outcome prediction. We chose and tested multiple candidate classifiers with different algorithms. Initially, we explored and compared the performance of the classifiers with the various algorithms (kNN, Random Forest, Decision Tree, and Logistic Regression). After testing the multiple candidate classifiers, we chose Random Forest, which yielded the best performance, as the main classification algorithm to perform model training and evaluation. We implemented 10-fold student-stratified cross-validation using the training dataset to evaluate the performance of each classifier.

Table 1. Ground-Truth Labelling Examples of Distress

Behaviour Examples
• After experiencing a game trial failure, a student suddenly lies back towards the chair back and sighs deeply.
• After experiencing a game trial failure, a student holds their chin and looks at the gameplay screen for a while.
• After experiencing a game trial failure, a student scratches their head and looks sombre.
• While playing a puzzle, a student slaps the laptop or desk.

Source: Biglan, 1991.

2.3. Performance Metrics

To select a main classifier for model development, we used various performance metrics results (i.e., accuracy, AUC, precision, recall, and F1 score). Table 2 demonstrates what the performance of each metric indicates. In addition, we also experimented with key hyperparameter settings of Random Forest. We tested a classifier's number of trees, pruning, and voting strategy to yield the best classification performance. We turned our hyperparameters through the training-fold dataset, not the test-fold one, because including test-fold data could inflate the performance estimates. Compared to traditional decision-tree algorithms, Random Forest uses ensemble learning for the number of random trees created on the bootstrapped subsets of datasets. Pruning reduces the size of decision trees by eliminating redundant and nonsignificant sub-trees in predicting a class. Pruning can also reduce model overfitting concerns. A voting strategy is a machine-learning model with ensemble learning, choosing a classification result based on the highest probability of a chosen class.

2.4. Feature Selection

2.4.1. Facial Expression Data

We used two different approaches to collect facial-expression data: action units (AUs) and image-based emotional probability. The existing FER data-driven emotion-recognition package (Giannopoulos et al., 2018; Goodfellow et al., 2013) provides public training data for emotion recognition to enable software developers to easily perform and integrate automatic emotion recognition in a computing system. However, the mere implementation of only a single open-source package may raise

concerns because the quality of the algorithms and training dataset in this package was not transparent or validated. Specifically, individual developers independently modify the algorithms and their features, so this may lead to low validity of emotion-recognition results. Therefore, using an additional independent facial recognition approach to validate its performance was necessary. Accordingly, we used OpenFace, designed to track facial-muscle movement recognition by collecting action unit (AU) data in a different way from that of the FER data-driven emotion-recognition package. In other words, the FER data-driven open-source package conducts emotion classification with ground-truth emotion labels based on the "big five" emotions (sad, angry, disgust, happy, and scared). On the other hand, AU-based software such as OpenFace only computes the intensity (r) of each AU-based muscle movement. AU intensity calculation involves a landmark detection technique that designates default positions of human facial muscles and checks the range of each muscle movement.

Table 2. Classifier Performance Metrics Used in this Study

Performance metrics	Description
Accuracy	The ratio of the number of accurate prediction results to the total number of input data classes. $$Accuracy = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}}$$
Area Under Curve (AUC)	Assesses the trade-off between true positive rate (recall) and false-positive rate. Used when dealing with a situation where the data sample distribution is skewed. $$AUC = \frac{S_0 - n_0\,(n_0 + 1)/2}{n_0 n_1}$$
Precision	Shows what percentage of the classification results are relevant. $$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$
Recall	The proportion of instances found to be relevant, which represents the algorithm's percentage of all relevant classification results. In other words, it indicates a classification's sensitivity. $$Recall = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$
F1 Score	The weighted mean of precision and recall scores. $$F1\ Score = \frac{2 \times (\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}}$$

Notes. n_0 and n_1 are the numbers of positive and negative examples respectively, and $S_0 = \sum r_i$, where ri is the rank of ith positive example in the ranked list.

2.4.2. Gameplay Logs

We gathered study participants' gameplay logs to track their gaming behaviours and performance. Both automatic and manual data collecting methods were used to obtain game logs. Specifically, gameplay efficiency and problem-solving phase data were collected via the manual codings of expert human coders. Game time duration, outcome, and game failure/success were automatically traced by a game system (one that embeds a logging mechanism to compute a list of features governing player actions). Gameplay logs include multiple gaming data features and denote how well each study participant completes game tasks in each gameplay round. For instance, gameplay logs depict how frequently study participants experience in-game success and failure during each trial of gameplay rounds. All coded features in the gameplay logs are presented in Table 3. To build a model, we utilized these game log-based data features that were deployed and reported in prior research (Asbell-Clarke et al., 2021; Liu & Israel, 2022; Rowe et al., 2021).

2.5. Feature Extraction

2.5.1. Dimensionality Reduction

We conducted dimensionality reduction to lower the number of input data features when building a prediction model. Dimensionality reduction is essential to building an efficient prediction model by trimming down unnecessary data features.

Otherwise, too many data features in the model training would lower the prediction quality. In other words, dimensionality reduction can eliminate model noise and enhance predictive accuracy with a simpler model. For dimensionality reduction, we conducted singular value decomposition (SVD; Mouri et al., 2019) in a current dataset. SVD reduces the data matrix to its components to simplify the calculation. We only included the data features for which all components had a cumulative variance greater than the threshold condition ($p > .95$). As a result, we included 36 data features (emotion probability, n = 5; OpenFace AUC, n = 17; gameplay log, n = 14) for the model development.

Table 3. Gameplay Log Data Features Used for Model Training

Gameplay log type	Subcategory	Definition	Literature
Game time duration		Gameplay time duration to complete a single game trial	Asbell-Clarke et al. (2021)
Game trial failure/success		Number of game trials (launching a mudball)	
Game outcome	Successfully moved all	Completely moved all available Zoombinis	Liu & Israel (2022) Rowe et al. (2021)
	Ran out of turns, moved some	Ran out of turns, moved some subset of available Zoombinis	
	All ejected	All Zoombinis ejected	
	Quit to map	Quit puzzle using the map button	
Gameplay efficiency level	Not at all efficient	A player throws mudballs in the same spot repeatedly; many empty cells are hit when there are better options	Rowe et al. (2021)
	Somewhat efficient	Although effective at times, a player does not take all known opportunities to connect dots; two or more errors	
	Highly efficient	A player takes advantage of nearly all known opportunities to connect the dots; only one error	
Problem-solving phases	Trial and errors	A player launches mudballs of varying colours and shapes at random	Liu & Israel (2022) Rowe et al. (2021)
	Systematic testing	A player demonstrates an organized, planned approach to testing hypotheses about the underlying row/column rules	
	Systematic testing with partial solution	After establishing the rule for one dimension (row or column), player systematically switches to testing the other dimension	
	Implementing a full solution	A player demonstrates deliberate targeting of grid cells with dots; rules for both rows and columns have been established	
Game strategy	Colour or shape constant	While systematically evaluating the rules of a row or column, player keeps the colour or shape constant; changes one attribute only between moves	Rowe et al. (2021)
	2D pattern completion	A player maintains a constant colour or shape while systematically testing shapes/colours to determine the rule of a row or column; A player also completes the entire row or column	
	Alternating colour/shape	To establish the rule of a row or column, A player alternates between keeping colour and shape constant	
	All combinations	A player experiments with all shape/colour pairs	
	Maximizing dots	A player actively targets dots	

Source: Asbell-Clarke et al., 2021.

2.6. Data Imputation

We carried out data imputation that replaces missing data with substitute data. Missing data is likely to cause data bias, limiting data analysis and interpretation. Data imputation is generally used to reduce data bias and preserve all data instances in a computational way (Jadhav et al., 2019). Missing data is especially likely to occur in multimodal data mining because it relies on inherent features of automatic data-tracking collection tools (e.g., facial emotion recognition). For example, if the automatic multimodal data collection causes computational errors when it does not have a particular training case for classification, missing data could occur, even in the same data channel. Once the data instance is missing in a single channel or multiple channels, all data instances involving the missing data should be excluded, which will subsequently influence the performance of a prediction model. Given that this study handles multimodal data that may include missing data, data processing to address this is particularly important for prediction-model development. Hence, we conducted data imputation using a k-nearest neighbour (kNN) algorithm (Malarvizhi & Thanamani, 2012). kNN imputes the missing data by using the value from the nearest record via a distance measure. Here, we set our k value at 5 and applied the Euclidean distance measure to the algorithm.

2.7. Synthetic Minority Oversampling (SMOTE)

We applied synthetic minority oversampling (SMOTE), a resampling technique that purposefully includes the minority cases in prediction model training (Fernández et al., 2018). This study aimed at predicting student distress states, so it required a sufficient set of data instances with ground-truth labelling results as "distress." However, with the current dataset captured during Zoombinis gameplay, the labelling results of distress were minority cases. In other words, the current dataset has an imbalanced data distribution, and the class with distress labels represents the minority. Imbalanced classification is a general issue in machine-learning prediction when class cases in the training dataset are not equally distributed. Imbalanced data causes poor predictive performance, specifically for the minority class. Therefore, we used SMOTE, which computationally copes with the imbalanced data by systematic resampling of the current training dataset. SMOTE selects one of the k nearest neighbours to a given data point at random, then generates a synthetic data point at a random place on the line between the provided data point and the chosen neighbour. SMOTE takes the kNN algorithm, by which we normalized the vector values to avoid a prediction problem caused by the magnitude of data values with different scales. We tested various numbers of neighbours in the kNN algorithm until we reached the best prediction performance.

Table 4. The List of Hyperparameters and Their Values in Classifier Development

Hyperparameter Type	Purpose	Details
Cross Validation	Model performance evaluation	10-fold cross validation Local random seed (n = 2000)
Stratified Sampling	Data sampling	Training (80%) and testing data (20%)
Singular Value Decomposition (SVD)	Dimensionality reduction	Removing the data features with the least impact under a threshold condition (threshold = 0.95)
Random Forest	Classification algorithm	Number of trees = 75 Criterion = Gain ratio Maximal depth = 10 Pruning (confidence = 10%) Voting = majority voting
k-Nearest Neighbours (kNN)	Data imputation algorithm	Number of k = 5 Mixed measures = Mixed Euclidean Distance
Synthetic minority Oversampling (SMOTE)	Data resampling technique for imbalanced data	Normalization Number of neighbours = 20 Nominal change rate = 50%

3. Results

3.1. Hyperparameter Settings

In terms of model performance evaluation, to reduce bias and maintain a modest level of data variance, we conducted 10-fold student-level cross validation where data points were divided into ten groups that belong to either the training or testing dataset for each fold. We applied stratified sampling that purposefully samples a test set that best represents the entire dataset. This sampling strategy is useful because it reduces the risk of sampling errors. We then adjusted several methodological features above, to refine the classifier settings: data imputation, dimensionality reduction, and SMOTE. Using kNN, we conducted data imputation to replace missing values in the prediction model's input data. Regarding the data features of the OpenFace-based classifier, we used the intensity (r) values of 18 AUs, whereas the classifier of FACE data-driven open source provided

probability estimates for each of the "big five" emotional states (sad, angry, disgust, happy, and scared). In addition, the classifier using gameplay logs included a total of 15 data features, such as game performance results, game strategies used, task efficiency, and gameplay time duration. Table 4 shows the hyperparameters and their settings for designing classifiers in this study.

3.2. Performance Results

Table 5 demonstrates the overall performance results of the unimodal classifiers (estimated emotion probabilities, action units, and gameplay logs) and the fused classifier. On average, all classifiers had good performance results (accuracy > .75, AUC > .80) when detecting the state of distress except for those from gameplay log data. Specifically, the fused classifier yielded the best accuracy — (p = .836), AUC performance (p = .850), and F1 score (p = .809) — in detecting the state of distress. Given the concern related to data imbalances in the current research, AUC, Recall, and F1 scores give a comprehensive result that indicates overall satisfactory model performance in sensitively detecting student distress. These results conclusively suggest that multimodal data fusion has improved tracking student distress states.

Table 5. Performance Results of Three Unimodal Classifiers and Fused Classifier

Performance metrics/Classifier	FER data-driven Open source Only	OpenFace Only	Gameplay Logs Only	Fused
Accuracy	.791	.770	.560	**.836**
AUC	.802	.805	.581	**.850**
Precision	.850	.771	.532	.844
Recall	.718	.772	.540	**.815**
F1 score	.778	.769	.698	**.809**

4. Discussion and Conclusion

4.1. Effects of Multimodal Data Fusion

We found that the classifier with decision-level multimodal data fusion outperformed the prediction by each of those classifiers with unimodal data sources. According to the performance results of machine-learning classifiers, the classifier with multimodal data fusion yielded better prediction performance than others. This study suggests that multimodal data fusion helps to improve the prediction of cognitive-affective states. Due to our interest in student multimodal interaction in immersive learning environments, this research has sought ways to collect and synthesize multimodal data for a comprehensive projection of student learning paths. Recent studies have tested multimodal data fusion to predict students' cognitive-affective states (Henderson et al., 2019; 2020) as well as learning achievement (Chango et al., 2021). The improved performance of the fused classifier in this study corroborates the usefulness of multimodal data fusion when building a learning analytics system.

The results of this study show that the classifier with multimodal data fusion yielded superior performance in most machine-learning metrics. It is noteworthy that the fused model generally outperformed the other models on the recall score, compared to that of other classifiers. A recall score relates to a prediction model's sensitivity, referring to the proportion of true-positive cases predicted as positive. It denotes a prediction model's actual performance in tracking a target cognitive-affective state. Since distress connotes game task-related unpleasant feelings, the accurate prediction of distress during educational gameplay can drive timely adaptive learning support. Moreover, the multimodal data fusion yielded the best AUC scores (.850) on tracking the states of distress. We speculate that multimodal data fusion better captures students' distress by synthesizing students' facial expressions and gameplay performance related to negative gaming results. An interpretation is that unimodal data sources from students' facial expressions or gameplay performance appear limited in tracking students' distress clearly, hence synthesizing these two types of data sources was helpful to corroborate the prediction on individuals' distress states.

4.2. Data Processing in Multimodal Data Fusion

This study demonstrates a comprehensive workflow for collecting, processing, and analyzing multimodal data to track students' cognitive-affective states during educational gameplay. In particular, the findings suggest the process and benefit of using multimodal game-based learning analytics for tracking cognitive-affective states related to students' learning curves. Here we collected different types of unimodal data (i.e., emotion probability, action unit, and gameplay performance) and computationally merged them to implement multimodal data fusion. This study provides an illustrative case of the key steps of data collection, data wrangling, and prediction model development, tailored to the nature of various multimodal data during educational gameplay.

At least two data processing steps were useful in building a multimodal data-driven prediction model of cognitive-affective states: 1) data imputation and 2) synthetic minority oversampling (SMOTE). First, data imputation helped researchers to build a model training set for multimodal data mining. Considering the high likelihood of erroneous recognitions in data-tracking software causing data loss, data imputation efficiently handles missing unimodal data. Second, SMOTE was useful in alleviating the data imbalance concern. Tracking certain cognitive-affective states may cause a data imbalance because proportionally collecting every ground-truth class of various cognitive-affective states is not feasible. Even in the context of educational gameplay, distress is likely to happen less than other cognitive-affective states, which necessitates using computational techniques to overcome data imbalance issues when training the model. SMOTE suggests a way to efficiently manage data imbalance when building a tracking model of cognitive-affective states in educational game research. The results of the study suggest the usefulness of data upsampling techniques in managing imbalanced data for multimodal data fusion (Henderson et al., 2020).

4.3. Adaptive Design of Immersive Learning

The current study findings have two implications for game-based learning analytics design. First, the study finding suggests that multimodal data fusion helps to track student distress states during gameplay. Distress may be associated with individual experiences with learning tasks. If student distress levels are overwhelmingly high, it may indicate that they may face challenges in completing game tasks. It may also suggest negative wheel-spinning processes, thus indicating a critical moment for adaptive delivery of learning support. In particular, the behavioural coding of the gameplay recordings suggests that study participants often underwent unproductive wheel-spinning experiences while experiencing distress states. As such, tracking student distress can be timely indicative of their negative wheel-spinning, which calls for real-time learning support. The study findings contribute to the literature by highlighting the feasibility of tracking distress as a potential indicator or trigger for adaptive, real-time learning supports in a digital learning setting. The current fused models with OpenFace and FER-2013 data has a short time span of affective state detection so it is feasible to be applied as real-time adaptive learning support tailored to individual distress states. Moreover, this study finding also contributes to existing research on detecting wheel-spinning experiences in digital learning environments. Previous research has primarily discussed that affective factors are closely associated with wheel-spinning experiences. They also focused on students' prerequisite learning performance and its efficiency as a strong predictor of wheel-spinning experiences (Beck & Rodrigo, 2014; Mu et al., 2020; Owen et al., 2019). Aligned with these findings, the results of the present study suggest that student distress states evidenced by their facial expression and gameplay performance could also serve a key indicator of wheel spinning states. Especially, the outperformed fused model, including both facial expression and gameplay performance, demonstrates the advantage of multimodal data fusion in predicting wheel-spinning. Future research will examine how various verbal and non-verbal data features in a game system can be selected, integrated, and deployed with multimodal data fusion to track student distress states as a proxy for wheel-spinning.

Second, the study finding confirms the feasibility of integrating learning analytics into an immersive, highly interactive learning environment. Immersive learning environments, such as digital games, generally require learners to use multiple sensor channels (verbal and non-verbal) during play. The study results demonstrate that multimodal data fusion for immersive learning environments is suitable because it can better capture learning-related behaviour cues since students usually express their affective states through multimodal behavioural cues. Using unimodal data collection (e.g., gameplay logging) is limited in gathering affective state evidence. The lower performance of unimodal data-driven models in the current study confirms that multimodal data fusion is more capable of providing a comprehensive data synthesis and diagnosis of individuals' learning states. For instance, including students' non-verbal behavior cues in immersive learning environments (e.g., gestures, prosodic features of voices, and eye-gaze patterns) with multimodal data fusion could help to detect students' task challenges. To extend this study, future research should include alternative sensory data to build effective multimodal data fusion models. Future studies could explore the applications of data fusion model features from this study to the other immersive learning contexts. Some data features (e.g., problem-solving phases, task outcome, and efficiency) with facial expression data from the present study could be fully applied to track cognitive-affective states in non-game immersive learning systems.

4.4. Limitations

This study has limitations. First, during the ground-truth labelling of distress, we relied on observing student facial expressions and gestures to determine distress classes. To validate the human-labelling results, future research should include additional data collection of students' self-reported reactions. Second, the current dataset appeared insufficient to build a generalizable prediction model beyond the game used. Hence, a larger multimodal dataset of more students' gameplay across game platforms is warranted for the further validation and improvement of the current distress prediction model. Third, this study did not involve students' speech data during gameplay. Although we reviewed students' discourses or utterances during gameplay to qualitatively gauge their distress state, we did not transcribe or include student speech data for multimodal data mining and

fusion. Given that students may express cognitive-affective states in speech, future studies should include the collection, wrangling, and analysis of the speech data for multimodal data fusion. Fourth, we only tested simplistic, supervised machine learning implementation of multimodal data fusion in this study. Further empirical research is warranted to test and compare the model performance of simplistic and weighted models to identify the optimized settings of multimodal data fusion in game-based learning research.

Declaration of Funding

This paper is based upon work supported by the U.S. Department of Education, Educational Innovative Research program, Award #U411C190179.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Alam, M. R., Bennamoun, M., Togneri, R., & Sohel, F. (2015). A confidence-based late fusion framework for audio-visual biometric identification. *Pattern Recognition Letters*, *52*, 65–71. https://doi.org/10.1016/j.patrec.2014.10.006

Asbell-Clarke, J., Rowe, E., Almeda, V., Edwards, T., Bardar, E., Gasca, S., Baker, R. S., & Scruggs, R. (2021). The development of students' computational thinking practices in elementary- and middle-school classes using the learning game, Zoombinis. *Computers in Human Behavior*, *115*, 106587. https://doi.org/10.1016/j.chb.2020.106587

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*(4), 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003

Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016, March). OpenFace: An open source facial behavior analysis toolkit. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1–10). IEEE. https://doi.org/10.1109/WACV.2016.7477553

Beck, J., & Rodrigo, M. M. T. (2014). Understanding wheel spinning in the context of affective factors. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th International Conference on Intelligent Tutoring Systems* (ITS 2014), 5–9 June 2014, Honolulu, HI, USA (pp. 162–167). Springer. https://doi.org/10.1007/978-3-319-07221-0_20

Biglan, A. (1991). Distressed behavior and its context. *The Behavior Analyst*, *14*(2), 157–169. https://doi.org/10.1007/BF03392566

Bosch, N., Chen, Y., & D'Mello, S. (2014). It's written on your face: Detecting affective states from facial expressions while learning computer programming. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th International Conference on Intelligent Tutoring Systems* (ITS 2014), 5–9 June 2014, Honolulu, HI, USA (pp. 39–44). Springer. https://doi.org/10.1007/978-3-319-07221-0_5

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L., & Zhao, W. (2015). Automatic detection of learning-centered affective states in the wild. *Proceedings of the 20th International Conference on Intelligent User Interfaces* (IUI '15), 29 March–1 April 2015, Atlanta, GA, USA (pp. 379–388). ACM Press. https://doi.org/10.1145/2678025.2701397

Bowman, N. D., & Tamborini, R. (2012). Task demand and mood repair: The intervention potential of computer games. *New Media & Society*, *14*(8), 1339–1357. https://doi.org/10.1177/146144481245042

Chrousos, G. P., Loriaux, D. L., & Gold, P. W. (2013). *Mechanisms of physical and emotional stress* (Vol. 245). Springer Science & Business Media.

Chango, W., Cerezo, R., & Romero, C. (2021). Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses. *Computers & Electrical Engineering*, *89*, 106908. https://doi.org/10.1016/j.compeleceng.2020.106908

Chen, F., Cui, Y., & Chu, M. W. (2020). Utilizing game analytics to inform and validate digital game-based assessment with evidence-centered game design: A case study. *International Journal of Artificial Intelligence in Education*, *30*(3), 481–503. https://doi.org/10.1007/s40593-020-00202-6

DeFalco, J. A., Rowe, J. P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B. W., Baker, R. S., & Lester, J. C. (2018). Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education*, *28*(2), 152–193. https://doi.org/10.1007/s40593-017-0152-1

Emerson, A., Cloude, E. B., Azevedo, R., & Lester, J. (2020). Multimodal learning analytics for game-based learning. *British Journal of Educational Technology*, *51*(5), 1505–1526. https://doi.org/10.1111/bjet.12992

Eseryel, D., Law, V., Ifenthaler, D., Ge, X., & Miller, R. (2014). An investigation of the interrelationships between motivation, engagement, and complex problem solving in game-based learning. *Journal of Educational Technology & Society*, *17*(1), 42–53. http://hdl.handle.net/20.500.11937/26368

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905. https://doi.org/10.1613/jair.1.11192

Filsecker, M., & Hickey, D. T. (2014). A multilevel analysis of the effects of external rewards on elementary students' motivation, engagement and learning in an educational game. *Computers & Education*, *75*, 136–148. https://doi.org/10.1016/j.compedu.2014.02.008

Fultz, J., Schaller, M., & Cialdini, R. B. (1988). Empathy, sadness, and distress: Three related but distinct vicarious affective responses to another's suffering. *Personality and Social Psychology Bulletin*, *14*(2), 312–325. https://doi.org/10.1177/0146167288142009

Giannopoulos, P., Perikos, I., & Hatzilygeroudis, I. (2018). Deep learning approaches for facial emotion recognition: A case study on FER-2013. In I. Hatzilygeroudis & V. Palade (Eds.), *Advances in hybridization of intelligent methods* (pp. 1–16). Springer. https://doi.org/10.1007/978-3-319-66790-4_1

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., ... & Bengio, Y. (2013, November). Challenges in representation learning: A report on three machine learning contests. ICML 2013 Workshop on Challenges in Representation Learning (pp. 117–124). Springer. https://doi.org/10.48550/arXiv.1307.0414

Gupta, H., Varshney, H., Sharma, T. K., Pachauri, N., & Verma, O. P. (2021). Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. *Complex & Intelligent Systems*, *8*, 3073–3087. https://doi.org/10.1007/s40747-021-00398-7

Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior*, *54*, 170–179. https://doi.org/10.1016/j.chb.2015.07.045

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 of the 20th International Conference on Artificial Intelligence in Education* (AIED 2019), 25–29 June 2019, Chicago, IL, USA (pp. 144–156). Springer. https://doi.org/10.1007/978-3-030-23204-7_13

Henderson, N., Rowe, J., Paquette, L., Baker, R. S., & Lester, J. (2020). Improving affect detection in game-based learning with multimodal data fusion. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, E. Millán (Eds.), *Proceedings of the 21st International Conference on Artificial Intelligence in Education* (AIED 2020), 6–10 July 2020, Ifrane, Morocco (pp. 228–239). Springer. https://doi.org/10.1007/978-3-030-52237-7_19

Hofmann, M., & Klinkenberg, R. (Eds.). (2016). RapidMiner: Data mining use cases and business analytics applications. CRC Press.

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, *33*(10), 913–933. https://doi.org/10.1080/08839514.2019.1637138

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L., Walters, E. E., & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, *32*(6), 959–976. https://doi.org/10.1017/s0033291702006074

LePine, J. A., LePine, M. A., & Jackson, C. L. (2004). Challenge and hindrance stress: Relationships with exhaustion, motivation to learn, and learning performance. *Journal of Applied Psychology*, *89*(5), 883–891. https://doi.org/10.1037/0021-9010.89.5.883

Liu, T., & Israel, M. (2022). Uncovering students' problem-solving processes in game-based learning environments. *Computers & Education, 182*, 104462. https://doi.org/10.1016/j.compedu.2022.104462

Malarvizhi, R., & Thanamani, A. S. (2012). K-nearest neighbor in missing data imputation. *International Journal of Engineering Research and Development*, *5*(1), 5–7. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.375.925&rep=rep1&type=pdf

Mouri, K., Suzuki, F., Shimada, A., Uosaki, N., Yin, C., Kaneko, K., & Ogata, H. (2019). Educational data mining for discovering hidden browsing patterns using non-negative matrix factorization. *Interactive Learning Environments*, *29*(7), 1176-1188. https://doi.org/10.1080/10494820.2019.1619594

Mu, T., Jetten, A., & Brunskill, E. (2020). Towards suggesting actionable interventions for wheel-spinning students. *Conference on Educational Data Mining* (EDM2020), 10–13 July 2020, Online (pp. 183–193). International Educational Data Mining Society. https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_201.pdf

Munshi, A., Rajendran, R., Ocumpaugh, J., Biswas, G., Baker, R. S., & Paquette, L. (2018, July). Modeling learners' cognitive and affective states to scaffold SRL in open-ended learning environments. *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization* (UMAP 2018), 8–11 July 2018, Singapore (pp. 131–138). ACM Press. https://doi.org/10.1145/3209219.3209241

Naylor, R. (2020). Key factors influencing psychological distress in university students: The effects of tertiary entrance scores. *Studies in Higher Education*, *47*(3), 630–642. https://doi.org/10.1080/03075079.2020.1776245

Ochoa, X., & Worsley, M. (2016). Augmenting learning analytics with multimodal sensory data. *Journal of Learning Analytics*, *3*(2), 213–219. https://doi.org/10.18608/jla.2016.32.10

Owen, V. E., Roy, M. H., Thai, K. P., Burnett, V., Jacobs, D., Keylor, E., & Baker, R. S. (2019). Detecting wheel-spinning and productive persistence in educational games. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.), *Proceedings of the 12th International Conference on Educational Data Mining* (EDM2019), 2–5 July 2019, Montréal, Quebec, Canada (pp. 378–383). International Educational Data Mining Society. https://files.eric.ed.gov/fulltext/ED599202.pdf

Plass, J. L., Homer, B. D., Pawar, S., Brenner, C., & MacNamara, A. P. (2019). The effect of adaptive difficulty adjustment on the effectiveness of a game to develop executive function skills for learners of different ages. *Cognitive Development*, *49*, 56–67. https://doi.org/10.1016/j.cogdev.2018.11.006

Rodrigo, M. M. T. (2011). Dynamics of student cognitive-affective transitions during a mathematics game. *Simulation & Gaming*, *42*(1), 85–99. https://doi.org/10.1177/1046878110361513

Rudland, J. R., Golding, C., & Wilkinson, T. J. (2020). The stress paradox: How stress can be good for learning. *Medical Education*, *54*(1), 40–45. https://doi.org/10.1111/medu.13830

Rowe, E., Almeda, M. V., Asbell-Clarke, J., Scruggs, R., Baker, R., Bardar, E., & Gasca, S. (2021). Assessing implicit computational thinking in Zoombinis puzzle gameplay. *Computers in Human Behavior*, *120*, 106707. https://par.nsf.gov/servlets/purl/10061932

Silvervarg, A., Haake, M., & Gulz, A. (2018). Perseverance is crucial for learning: "OK! but can I take a break?" In C. P. Rosé et al. (Eds.), *Proceedings of the 19th International Conference on Artificial Intelligence in Education* (AIED 2018), 27–30 June 2018, London, UK (pp. 532–544). Lecture Notes in Computer Science vol. 10947. Springer. https://doi.org/10.1007/978-3-319-93843-1_39

Shute, V. J., D'Mello, S., Baker, R., Cho, K., Bosch, N., Ocumpaugh, J., Ventura, M., & Almeda, V. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education*, *86*, 224–235. https://doi.org/10.1016/j.compedu.2015.08.001

Taub, M., Azevedo, R., Bradbury, A. E., & Mudrick, N. V. (2020). Self-regulation and reflection during game-based learning. In J. L. Plass, R. E. Mayer, & B. D. Homer (Eds.), *Handbook of game-based learning* (pp. 239–262). The MIT Press.

Verma, V., Rheem, H., Amresh, A., Craig, S. D., & Bansal, A. (2020, December). Predicting real-time affective states by modeling facial emotions captured during educational video game play. In I. Marfisi-Schottman, F. Bellotti, L. Hamon, & R. Klemke (Eds.), *Proceedings of the International Conference on Games and Learning Alliance* (GALA 2020), 1–3 December 2020, La Spezia, Italy. Lecture Notes in Computer Science vol. 12517. (pp. 447–452). Springer. https://doi.org/10.1007/978-3-030-63464-3_45

Yeh, Y. C., Lai, G. J., Lin, C. F., Lin, C. W., & Sun, H. C. (2015). How stress influences creativity in game-based situations: Analysis of stress hormones, negative emotions, and working memory. *Computers & Education*, *81*, 143–153. https://doi.org/10.1016/j.compedu.2014.09.011