Predictive Student Modeling in Game-Based Learning Environments with Word Embedding Representations of Reflection

Michael Geden, Andrew Emerson, Dan Carpenter, Jonathan Rowe, Roger Azevedo & James Lester

International Journal of Artificial Intelligence in Education
Official Journal of the International AIED Society

ISSN 1560-4292
Volume 31
Number 1

Int J Artif Intell Educ (2021) 31:1-23
DOI 10.1007/s40593-020-00220-4

Springer

International Journal of Artificial Intelligence in Education (2021) 31:1–23
https://doi.org/10.1007/s40593-020-00220-4

ARTICLE

Predictive Student Modeling in Game-Based Learning Environments with Word Embedding Representations of Reflection

Michael Geden[1] · Andrew Emerson[1] · Dan Carpenter[1] · Jonathan Rowe[1] · Roger Azevedo[2] · James Lester[1]

Accepted: 23 September 2020 / Published online: 20 October 2020

Abstract

Game-based learning environments are designed to provide effective and engaging learning experiences for students. Predictive student models use trace data extracted from students' in-game learning behaviors to unobtrusively generate early assessments of student knowledge and skills, equipping game-based learning environments with the capacity to anticipate student outcomes and proactively deliver adaptive scaffolding or notify instructors. Reflection is a key component of self-regulated learning, and it is critical in effective learning. However, there is currently limited work exploring the utility of reflection for inducing accurate predictive student models. This article presents a predictive student modeling framework that leverages natural language responses to in-game reflection prompts to predict student learning outcomes in a game-based learning environment for middle school microbiology, CRYSTAL ISLAND. With data from a pair of classroom studies involving 118 middle school students, we investigate the accuracy of early prediction models that utilize features extracted from student trace data combined with word embedding-based representations (i.e., GloVe, ELMo) of student reflection responses. We evaluate the accuracy of the predictive models over time using data from incremental segments of each student's interaction with the game-based learning environment, and we compare against models that omit student reflection features. Results reveal that models encoding students' natural language reflections with ELMo word embeddings yield significantly improved accuracy compared to other representations, with the greatest accuracy demonstrated by an ensemble of predictive models. We discuss the implications of these results for the design of game-based learning environments.

Keywords Student modeling · Early prediction · Game-based learning environments · Self-regulated learning · Reflection

✉ Andrew Emerson
ajemerso@ncsu.edu

Extended author information available on the last page of the article

Introduction

Game-based learning holds considerable potential for student engagement, learning, and motivation (de Freitas 2018; Plass et al. 2020; Qian and Clark 2016). Because game-based learning environments can create learning experiences that are effective and engaging, they have shown success across a diverse range of subject matter (Brezovszky et al. 2019; Pesare et al. 2016; Tsatsou et al. 2019). Game-based learning environments can provide students with adaptive support in the form of personalized hints and feedback as well as tailored curricular sequencing. Delivering adaptive support requires access to a student model to estimate the learner's current knowledge and competencies that can drive pedagogical decision making and timely interventions. Accurate student models are critical in successfully employing adaptive support, as incorrectly tailored support can negatively affect student learning outcomes (San Pedro et al. 2013; Chen and Law 2016).

There are multiple approaches to student modeling, including stealth assessment and knowledge tracing, which aim to model student competencies based on their interactions with a learning environment. These approaches, while exhibiting significant differences, both require an explicit mapping from student actions to knowledge content that is developed by subject matter experts. This mapping can be challenging to generate for game-based learning environments, where an action may not neatly pair with specific content knowledge or skills (e.g., movement, picking up item). Another important approach to student modeling is to predict future student competencies using data acquired from a student up to a specific moment. This type of modeling, *predictive student modeling,* automatically induces a student model using features extracted from trace logs of students' gameplay interactions to predict student learning outcomes, eliminating the need for labor-intensive encoding of human domain knowledge. Recent years have seen significant interest in machine learning approaches to predictive student modeling in game-based learning environments (Akram et al. 2018; Emerson et al. 2019; Falakmasir et al. 2016; Karumbaiah et al. 2018; Min et al. 2017b, 2020; Sabourin et al. 2013b). Deep neural networks, in particular, have demonstrated significant improvements over other machine learning approaches for inducing accurate predictive student models (Akram et al. 2018; Geden et al. 2020; Min et al. 2017b, 2020; Piech et al. 2015; Wang et al. 2017), although these improvements have not proven universal (Xiong et al. 2016).

A promising direction for enriching predictive student models is to leverage evidence of self-regulated learning (SRL). Effective use of SRL strategies has been shown to yield positive learning outcomes in game-based learning environments (Nietfeld et al. 2014; Sabourin et al. 2013a; Taub et al. in press). Predictive student models that leverage trace log data interactions have been shown to model SRL behaviors in a wide range of computer-based learning environments (Biswas et al. 2018; Desmarais and Baker 2012; Kinnebrew et al. 2015). Reflection is a key component of SRL (Greene and Azevedo 2007; Schunk and Greene 2018; Winne and Hadwin 1998) that involves the deliberate examination of one's own knowledge, skills, motivational beliefs, emotions, and goals following a learning activity (e.g., post-problem-solving episode). A common approach to examining students' reflection processes is prompting students to create written (i.e., textual)

descriptions of their reflections (Thorpe 2004; Minott 2008; Hume 2009; Bannert and Reimann 2012; Van den Boom et al. 2004). Reflection responses may describe new information that the student has learned or goals that the student may have recently accomplished or a new understanding of potentially novel ways to learn, use strategies, and acquire skills. Reflection can be prompted periodically during a student's interaction with a game-based learning environment, or it can be prompted after a student has finished interacting with the game.

To assess students' reflection processes, written reflections are often annotated using a manual coding process. Manual coding is labor-intensive (Shute 2011; Poldner et al. 2014; Prilla and Renner 2014; Bannert and Reimann 2012), which is an obstacle to providing real-time adaptive support for student reflection. An alternative approach is utilizing natural language processing to automatically assess students' written reflections. Automated assessment methods have been devised that employ latent semantic analysis (LSA) to derive features from text responses to predict the quality of student reflections (Cheng 2017; Kovanović et al. 2018). Word embedding techniques have also been investigated (Pennington et al. 2014; Peters et al. 2018). Word embedding techniques enable automatic distillation of syntactic and semantic meaning from text, and they have been shown to consistently outperform LSA with respect to capturing word relationships. Using these automated text analysis methods, it is possible to build predictive student models that utilize features from written reflections to drive real-time adaptive support in game-based learning environments.

In this work, we incorporate student written responses to in-game reflection prompts using word embedding methods for early prediction of student learning. We extract features from students' reflections in an unsupervised fashion by using GloVe and ELMo pre-trained word embeddings (Pennington et al. 2014; Peters et al. 2018), which have been shown to yield significant benefits in automated text analytics. GloVe word embeddings are generated using a co-occurrence matrix derived from a large corpus of text. ELMo embeddings are generated with stacked bi-directional LSTMs and capture additional information related to the specific context in which a word was used. By using word embeddings that are pre-trained on a large text corpus, we are able to capture complex syntactic and semantic characteristics of students' natural language reflections from a comparatively small dataset. We compare predictive models that use word embedding representations to models that do not incorporate student reflection features using a range of supervised learning techniques, including deep recurrent neural networks, ensemble techniques, and several traditional classification algorithms (e.g., naïve Bayes, k-nearest neighbor, support vector machines). We explore the effectiveness of this approach in a game-based learning environment for middle school microbiology, CRYSTAL ISLAND. Word embedding-based reflection feature representations are evaluated on a dataset consisting of log data from student interactions with CRYSTAL ISLAND during two classroom studies ($N = 118$). In addition to evaluating model accuracy, we measure the success of early prediction models by evaluating the standardized convergence point and convergence rate of each model (Min et al. 2020; Blaylock and Allen 2003; Min et al. 2016). Results point toward the promise of leveraging features extracted from students' reflection processes to devise predictive student models that can inform adaptive support in game-based learning environments.

Background

Reflection

Reflection is a key component of self-regulated learning (Greene and Azevedo 2007; Winne and Hadwin 1998; Winne and Azevedo 2014). Reflection is a deliberate process that involves performing an introspective examination of one's own experience, perspective, and feelings with respect to a personally relevant problem (Ullmann 2015). Effective reflection can lead students toward novel insights about their own learning (Rogers 2001), resulting in novel understanding, new perspectives, or the acquisition of knowledge and skills (Schön 2017; Boud et al. 2013). Self-regulatory strategies, such as reflection, are critical to student learning (Greene et al. 2011) and can lead to positive learning outcomes in adaptive learning environments (Van den Boom et al. 2004; Bannert 2006; Van den Boom et al. 2007).

A common practice for fostering reflection is prompting students to engage in written reflection after a learning activity (Thorpe 2004; Minott 2008; Hume 2009). Reflective writing can enhance learning outcomes by encouraging students to think critically about the educational materials and their meta-cognitive judgments (Bisman 2011; Burrows et al. 2001; Cisero 2006). There are many formats for reflection prompts depending upon the context in which they are used, but generally, they entail individuals writing about their thoughts and feelings on their previous experiences. While written reflections produce a measurable artifact of an individual's reflection process, assessing student reflection typically requires labor-intensive manual coding (Shute 2011; Poldner et al. 2014; Prilla and Renner 2014).

There are numerous methods for assessing reflection quality, but they broadly belong to two types of measurement: depth and breadth. Depth of reflection represents the sophistication and complexity of a student's reflection, and it is typically measured on a binary (e.g., non-reflective, highly reflective; Lai and Calandra 2010) or ordinal scale (Kember 1999; Poldner et al. 2014; Wong et al. 1995). Ordinal scales of reflection can involve different levels of reflection complexity (Wise and Jung 2019), forms of reflection (Lai and Calandra 2010), or types of information included in the reflection (Poldner et al. 2014). Breadth of reflection can be assessed by categorizing the range of subjects present in written reflections, such as understanding, content, assumptions, perspectives, and personal beliefs (Prilla and Renner 2014; Kember et al. 2008; Boenink et al. 2004; Ullmann 2019; Bell et al. 2011).

There is a growing interest in developing and utilizing automated methods for assessing reflection by analyzing written responses to reflection prompts. While there has been some work on modeling students' reflective process as a whole (Blaney et al. 2014; Ullmann 2011, 2019), the majority of existing work has investigated the primary subjects of students' reflections, such as their experiences, difficulties, intentions, and beliefs (Cheng 2017; Kovanović et al. 2018). Prior work has largely utilized dictionary methods and latent semantic analysis to extract features for training machine learning models such as naïve Bayes, support vector machines, or random forests to predict metrics of student reflection (e.g., quality, category). We utilize word embedding techniques to extract general syntactic and semantic features from the reflective text in lieu of latent semantic analysis due to their ability to transfer accurate representations leveraged from large corpora (Peters et al. 2018). We then provide the word embedding-based representation of student reflection directly to a predictive student model.

Predictive Student Modeling

Devising predictive models of student knowledge and behavior in game-based learning environments has shown significant promise (Min et al. 2016, 2017a; Wang et al. 2017; Emerson et al. 2019; Wu et al. 2019). *Predictive student modeling* focuses on making early predictions about students' future performance, knowledge, or skills using currently available data. A primary characteristic of predictive student modeling is that predictive student modeling infers future performance rather than current knowledge and skills. For example, knowledge tracing typically focuses on modeling whether a student has mastered a particular knowledge component or skill during their interactions with an adaptive learning environment. Knowledge tracing is often performed using hidden Markov models or recurrent neural networks (Baker et al. 2008; Piech et al. 2015). Stealth assessment uses trace data from student interactions with a game-based learning environment to predict student knowledge based on an evidence-centered design (ECD) framework (Shute et al. 2016; Kim et al. 2016). Computational models of stealth assessment often utilize Bayesian networks to link individual student actions to content knowledge. Temporal representations of student data can be captured through the use of sequence-based models, such as long short-term memory (LSTM) networks. In Min et al. (2020), LSTMs were used to predict student post-test performance following interactions with a game-based learning environment for middle grade computational thinking. This approach, called *deep stealth assessment*, minimizes the need to manually create a mapping between in-game actions and student knowledge; instead, the mapping is constructed automatically using deep learning techniques. In our work, we utilize a similar formulation of student modeling and develop predictive models that converge toward an accurate early prediction of student post-test performance, assessing model fit through a diverse set of accuracy and convergence metrics. Additionally, we investigate the use of students' in-game reflection responses as a complementary data channel with which to model post-test performance. While previous work has investigated using data from student dialogue with virtual agents in adaptive learning environments to inform models of student learning (Graesser 2016), utilizing student responses to in-game reflection prompts to enhance predictive student models is a novel contribution of this work.

Early Prediction

Early prediction enables adaptive learning environments to proactively deliver support to students (Acharya and Sinha 2014; Mao et al. 2019; Sabourin et al. 2013b; Winchell et al. 2018; Emerson et al. 2019). There is a growing literature on methods for accurately predicting student performance at key early points in a student's interaction with an adaptive learning environment. Common predictor features include static attributes such as demographic data, survey data, and pre-test scores. Olivé et al. (2019) use static features extracted from student data collected up to two days before an assignment deadline to predict on-time submissions. Other feature representations include sequential data on students' academic behavior or interactions with an adaptive learning environment. For example, Jiménez et al. (2019) use sequences of student data (e.g., course enrollment history, grades) in a three-year computer science program with the goal of finding the earliest moment in a student's academic career in which a reliable prediction of dropout is possible.

Many early prediction tasks in adaptive learning environments are formalized as classification problems. A classic example is predicting whether a student will be a high or low performer in a course (Costa et al. 2017; Gitinabard et al. 2019; Polyzou and Karypis 2019). Student performance is often discretized by performing a median split on student outcome data or by distinguishing between pass versus fail outcomes. Gitinabard et al. (2019) predicted student performance in a blended class that mixed in-person instruction with online learning platforms, splitting students into two groups: A- and above, and B+ and below. They segmented the data by defining a time threshold for consecutive actions in the online platforms. Then, they extracted features from the segmented data at different points in the semester, identified the most predictive features, and compared the predictive performance of using features encompassing the entire semester to features extracted using only data from earlier points in the semester. Polyzou and Karypis (2019) predicted student course grades in future semesters, with a focus on poor-performing students.

When making early predictions about student outcomes, it is common to use a small set of discrete data segments defined by course milestones (e.g., semester start, semester end, assignment due dates, examination dates). For example, Baneres et al. (2019) split a course timeline into deciles to generate predictions about student success at each time point. Kostopoulos et al. (2019) split the academic year into four consecutive time periods, separating demographic and survey data from time-variant data as two "views" in order to predict if a student will pass or fail a class. These methods require access to the entire set of data prior to segmentation and would be challenging to apply to game-based learning environments where there can be an absence of clear event boundaries. A range of computational techniques have been utilized for early prediction, with recent years focusing increasingly on deep neural networks (Lykourentzou et al. 2009; Min et al. 2020). Botelho et al. (2019) built multi-task deep learning models to predict student attrition and wheel spinning at each learning opportunity in an adaptive learning environment.

We build upon the literature on predictive student modeling by leveraging deep recurrent neural networks to generate early predictions of student post-test performance following interactions with a game-based learning environment. We show how the predictive model performs during early phases of students' action sequences, and we use an absolute measure of time (i.e., minutes elapsed) rather than a relative measure (e.g., percentage-based) to segment student data. Furthermore, we utilize features extracted from students' in-game reflection responses, yielding an enriched view of student learning processes in order to enhance the predictive accuracy of student models in game-based learning environments.

Crystal Island

CRYSTAL ISLAND is a game-based learning environment for middle school microbiology education (see Fig. 1). Students take the role of a medical investigator who has recently arrived on a remote island research station to identify the cause of an outbreak that is spreading among a team of scientists. Students must identify what type of pathogen is causing the illness, what is the disease's transmission source, and what is the proper treatment for the disease. The events of CRYSTAL ISLAND take place in an open-world game environment. Students are free to explore the island and interact with any of the island's non-player characters (NPCs) or educational resources. Students can ask NPCs about their symptoms, read in-game informational texts about

Fig. 1 CRYSTAL ISLAND game-based learning environment

various microorganisms and illnesses, and test various food objects for pathogenic contaminants in a virtual laboratory. Additionally, students fill out a diagnosis worksheet to keep track of key information they have learned during their investigation. CRYSTAL ISLAND has been used by thousands of middle school students worldwide, and it has been shown to yield significant science learning benefits (Lester et al. 2014; Meluso et al. 2012; Rowe et al. 2011).

Students interacted with a version of CRYSTAL ISLAND that periodically prompted them to reflect on what they have learned and their upcoming plans in the game (i.e., "In your own words, please describe the most important things that you've learned so far, and what is your plan moving forward?"). The embedded reflection prompts were designed to elicit reflective thinking, thus encouraging students to monitor and adapt their self-regulated learning processes based on their previous game-based learning behavior and problem-solving plans (van den Boom et al. 2007). After the reflection prompt, students were instructed to self-report their problem-solving progress on a 10-point Likert scale. The prompts were delivered following key milestones in the game's science problem scenario.

Students were prompted to reflect on their actions thus far and to comment on their plans moving forward (Table 1). The reflection prompt took the form of a message, delivered via the student's in-game smartphone, requesting an update about the student's investigation (see Fig. 2). When students finished the game, they were asked to reflect on whether their approach to solving the mystery was successful, and whether they would do anything differently when solving a similar problem in the future. In total, students spent an average of about 6 min interacting with reflection prompts, corresponding to approximately 8% of their total time in CRYSTAL ISLAND.

Student responses to the reflection prompts varied significantly in length, specificity, quality, and content (Table 2). Some responses exhibited limited reflection by the student, providing only proximal information about the student's latest and current actions. (See second row of Table 2.) Other students provided detailed responses which incorporated updates about their beliefs and hypotheses. (See fourth row of Table 2).

Methods

Participants

Eighth-grade students from a middle school in the mid-Atlantic region were enrolled in 2018 and 2019 as part of two separate classroom studies. The studies shared the same

Table 1 In-game reflection prompts

Trigger	In-game Investigation Update Request	Presence of Progress Self-report
After talking to the camp nurse for the first time	*"Agent, it looks like you've spoken with the camp nurse. Before you continue, we'd like a report on your progress"*	Yes
After reading six virtual books	*"Agent, it looks like you've found several materials that may be useful. Before you continue, we'd like a report on your progress."*	Yes
After obtaining a positive test result in the virtual laboratory	*"Agent, it looks like you found an object that tested positive for pathogenic contaminants. Before you continue, we'd like a report on your progress."*	Yes
After submitting a proposed diagnosis and getting it wrong.	*"Agent, it looks like you are making progress on diagnosing the illness, but you're not quite there yet."*	Yes
After solving the mystery	*"Well done, Agent! You've saved everyone on the island. Now that you are finished, we would like to ask a couple of final questions. "* - *"Please explain how you approached solving the mystery".* - *"If you were asked to solve a similar problem in the future, what would you do the same and/or differently?"*	No
After time expires, but the student has not solved the mystery	*"Thank you for playing CRYSTAL ISLAND. Now that you are finished, we would like to ask a couple of final questions."* - *"Please explain how you approached solving the mystery."* - *"If you were asked to solve a similar problem in the future, what would you do the same and/or differently?"*	No

version of CRYSTAL ISLAND, procedure, and materials, but there were also several differences. The studies were conducted in different classrooms and at different points in the semester. We combine the data collected across the 2018 ($n = 61$) and 2019 ($n = 95$) studies into a single dataset and analyze it in aggregate. A binary variable representing which data collection the student belonged to was added as a feature to

Fig. 2 In-game reflection prompt

Table 2 Examples of reflection responses

1	*"I have learned that the disease is either a mutagen, pathogen, or carcinogen. Moving forward, I am going to discover which one the disease on the island is."*
2	*"I think I have a diagnosis. I am going to test my findings."*
3	*"To move forward I will just keep testing foods."*
4	*"I learned that the disease spreading is either a pathogen, mutagen, or carcinogen, but I am leaning towards pathogen because of how I think it spread. I am talking to the patients now to learn their symptoms in order to verify."*
5	*"The sickness is passing from one person to the other"*

account for systematic differences between years. After removing students with missing data, the final dataset was composed of data from 118 students. Missing data was due to a number of factors, including student absences, failure to complete the post-study surveys, and forgetting assigned participant IDs during the pre-survey or post-survey phases of the study. There were 37 participants with missing pre- or post-survey data. The average age of students was 13.6 ($SD = 0.5$). There was approximately an equivalent number of students who identified as male ($n = 55$) and female ($n = 60$), with 3 students responding as Other. There were 43 students who identified as White, 32 as Black or African American, 21 as Hispanic or Latino, 3 as Asian, 1 as American Indian or Alaskan Native, and 18 as Other.

Measures

Demographics The pre-study survey began with a short series of demographic questions involving the student's current age, gender, and race. None of these items were included as features in the predictive student model to avoid disparate treatment of students based on protected attributes.

Emotions and Values Questionnaire Following the demographic questions, the pre-study survey contained a 16-item questionnaire on emotions using a 5-point Likert scale adapted from Pekrun et al. (2011). The response options pertained to the student's degree of agreement about their contemporaneous experience of the following emotions: happy, hopeful, proud, angry/frustrated, anxious/fearful, ashamed, hopeless, bored, surprised, contemptuous/disgusted, neutral, confused, interested/curious, sad, eureka/sudden understanding, and that the task/activity is valuable. The items were treated as continuous features for model input.

Microbiology Content Knowledge Test The pre-study survey included a 17-item multiple-choice content knowledge assessment on the student's microbiology content knowledge ($M = 6.78$, $SD = 2.75$). This pre-test consisted of thirteen factual and four application questions. An example question from the test was "Which of the following diseases is caused by a viral infection?" The response options included, "Anthrax," "Smallpox," "Salmonellosis," and "Sickle Cell." The items were converted to binary variables based on the correctness of the student's responses, and they were included in the predictive model as input features. After the student played CRYSTAL ISLAND, they were presented with a post-test that contained a separate set of 17-items addressing

the same microbiology content knowledge ($M = 7.36$, $SD = 3.36$). The post-test consisted of fourteen factual and three application questions. All of the information needed to answer the questions on both assessments could be found in CRYSTAL ISLAND, whether through books or articles, conversations with in-game characters, or by viewing informational posters. Pearson correlations indicated significant relationships between the pre- and post-tests ($r = 0.53$, $p < 0.01$). Both the pre- and post-tests were created by an interdisciplinary team of researchers.

Gameplay Features While students played CRYSTAL ISLAND, the game automatically time-stamped and logged their gameplay actions. Similar to Min et al. (2017a), we encoded students' in-game actions in terms of several components. The first component was the completion of any of the eight key milestones within the game (e.g., talked to the camp nurse, submitted a diagnosis for the epidemic). The second component was the action type, which denoted whether the student had moved to a new location, conversed with an NPC, read a virtual book or article, completed a problem-solving milestone (e.g., submitting a correct diagnosis worksheet), used the testing equipment in the virtual laboratory, or recorded notes in the diagnosis worksheet. In this study, the data included nine action types. The third component of an in-game action was the action argument, which varied based upon the action type. Action arguments included the titles of virtual books, the names of NPCs with whom the student conversed, the type of food object the student tested in the virtual laboratory, and the content entered into the diagnosis worksheet. The data included 95 different action arguments. The fourth component of an in-game action encoding was the location where the action occurred. The data included 24 non-overlapping, discrete locations in the virtual environment. Finally, three of the features were count-based action arguments that pertained to only a specific action type (e.g., number of edits in the diagnosis worksheet). In total, this yielded 130 distinct action features (8 milestones, 24 locations, 95 binary actions, 3 count-based actions). We represent actions with a one-hot encoding of each action component (i.e., action type, action argument, action location) to produce 127 binary features and three count-based features. Each encoded action also included the timestamp for the action's occurrence during a gameplay session.

Reflection Prompt Responses At several milestones in the game, students were prompted to reflect on their learning by providing a free-response description of their recent progress and upcoming problem-solving plans. Students also responded to a self-report question about their progress using a 10-point Likert scale ($M = 6.6$, $SD = 2.3$). Students were prompted up to five times at selected trigger points in the game chosen to minimize disruption to gameplay. (See the left column in Table 1 for several example trigger points.) Students' reflections added information to the predictive student models that was complementary to the pre-test and gameplay data to describe student behavior.

Procedure

After students completed the pre-study survey, researchers briefly described the overall purpose of the study. Next, all students watched a brief video that introduced the game's science mystery storyline. Afterward, students began

playing the tutorial phase of the game, which introduced them to the controls and game environment. Once finished with the tutorial, students had full agency to explore the island and investigate the mystery. Students in the 2018 study were allotted approximately two class periods and students in the 2019 study were allotted approximately three class periods to play the game until they either solved the mystery or ran out of time. The average playtime across all students was 76.3 min ($SD = 19.5$). After all students had either solved the mystery (72%) or run out of time (28%), they were directed to the post-study survey.

Reflection Writing Processing

Students' written responses to the reflection prompts underwent a series of pre-processing steps. First, punctuation was removed and all text was converted to lowercase. Next, misspellings were identified and corrected with a plausible candidate that was found within a Levenshtein (edit) distance of two. Incorrectly spelled words for which a plausible correction could not be found were removed. Word embeddings for each response were computed using one of two methods: 300-dimensional GloVe embeddings (Pennington et al. 2014) or 1024-dimensional ELMo embeddings (Peters et al. 2018). These embedding sizes were chosen because they are the most similar in size—the largest pre-trained GloVe embeddings are of size 300, and the smallest pre-trained ELMo embeddings are of size 1024. GloVe generates word embeddings based on non-sparse elements of a global co-occurrence matrix through a log-bilinear regression on a local context window. ELMo generates context specific word embeddings through concatenating the output of the top layer of stacked bi-directional LSTMs. We utilized the final hidden layer of the ELMo language model to derive embeddings.

A single representation of each reflection response was then created by averaging the word-level embeddings (Wieting et al. 2016; Adi et al. 2017). Word embeddings were averaged along each dimension across all words included in a student's written reflection. The limited sample size made the averaged word embedding approach advantageous compared to a sequential representation of the word embeddings, because a sequential approach (e.g., GRU, LSTM) would require learning many more model parameters. Additionally, averaging word embeddings has been shown to perform similarly or better than more complex methods across a number of datasets (Shen et al. 2018). These average word embeddings differentiated between reflection responses based on the words used and the structure of the response. For example, responses that included statements that were structurally similar to "I have learned...disease/illness...I plan to…" had similar embeddings, as did responses that contained phrases like "tested positive for pathogenic/nonpathogenic bacteria/virus" (see Fig. 3). Finally, a single representation of the student's overall reflections at the time of model prediction was created by averaging their mean reflection embeddings across all reflection prompts to which they had responded. For example, if the student, at the 16-min mark, had responded to two reflection prompts, then their reflection features would be the average of their two word-level averaged embeddings up to that point. At time points prior to the student's first reflection prompt, a zero-vector was used.

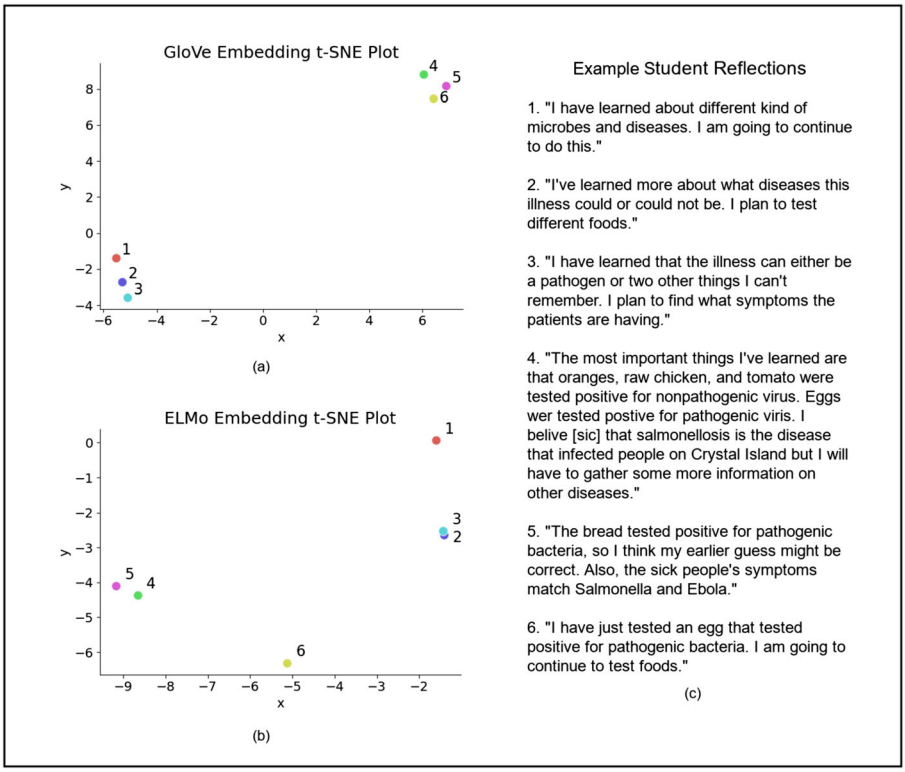

Fig. 3 t-SNE visualizations of embedding-based representations of student reflections: (a) GloVe embedding t-SNE plot, (b) ELMo embedding t-SNE plot, (c) Example student reflections

Results

All models were constructed in Python using Keras and Scikit-learn (Pedregosa et al. 2011). Model training and evaluation were conducted using 10-fold cross-validation on the student-level. The outcome variable was a binary indicator derived from a median split of post-test scores (low/high performance). The median post-test score was 7, and the split exhibited an approximate bimodal distribution (see Fig. 4). By using this outcome variable, the predictive student modeling task centered on predicting student

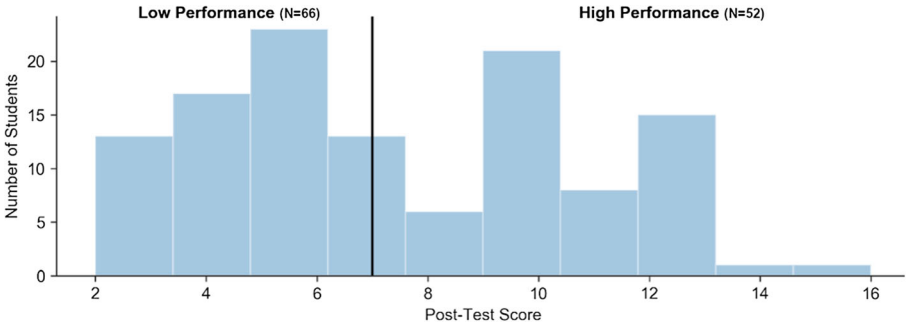

Fig. 4 Histogram of student post-test scores

knowledge at the end of the session, rather than predicting the change in knowledge during the session. The input features for the models consisted of each question from the emotions questionnaire ($k = 16$), pre-test questions ($k = 17$), and a one-hot encoding of student gameplay actions ($k = 131$). The predictive accuracy of the models was compared across three conditions, which varied in their utilization of reflection feature data: no reflection data, GloVe embeddings of reflection data ($k = 300$), and ELMo embeddings of reflection data ($k = 1024$). The conditions incorporating the reflection embeddings also included the self-report Likert item asking about the student's problem-solving progress.

Student performance on the post-test was predicted at two-minute intervals of gameplay as well as at the conclusion of the game. Given that large sections of the game involve reading text, this time interval allows for the inclusion of several game actions and events within each prediction interval. For each interval, we used a cumulative representation of student data up to that point. For example, the first interval's prediction was based upon the first two minutes of gameplay for each student. The second interval's prediction was based upon the first four minutes, and so on. The construction of the gameplay intervals caused students with long gameplay durations to be overrepresented during training compared to students who solved the problem scenario quickly. To counterbalance this effect, within each training fold, the training data was sampled with replacement so that all students had the same number of instances. Early prediction was evaluated using standardized convergence point with no penalty (Min et al. 2016) and convergence rate (Blaylock and Allen 2003), as well as several standard classification metrics (i.e., F1 score, precision, recall, AUC, and accuracy). Standardized convergence point is the proportion of the sequence of predictions which had converged, with 0 indicating the sequence never converged and 1 indicating an entirely accurate sequence. Convergence rate is the proportion of sequences with an accurate final prediction when utilizing the entirety of the student's data.

Several classification algorithms were compared across each of the embedding conditions (no reflection features, GloVe embeddings of reflection, ELMo embeddings of reflection): majority classifier, naïve Bayes, k-nearest neighbors (KNN), linear support vector machine (LinearSVM), logistic least absolute shrinkage and selection operator (LASSO), random forest, and Adaboost. The input features were the pre-test items, the sum of the one-hot encoded gameplay features, elapsed game time, and the reflection embeddings. All features were standardized within the 10-fold cross-validation, and then hyperparameter tuning of non-neural classification models occurred on an additional internal 10-fold cross-validation. The internal cross-validation split the training set into a training and validation set which were iteratively used to evaluate a range of model hyperparameters.

In addition, a recurrent neural network was evaluated that utilized a sequential representation of the gameplay features. The neural network consisted of a gated recurrent unit (GRU; Cho et al. 2014) for the gameplay features and a fully connected layer for the concatenation of the pre-test and reflection features. The output of the GRU and fully connected layers were then concatenated and propagated into an additional fully connected layer before the output layer (Fig. 5).

For regularization, dropout was applied after the GRU and both fully connected layers at a rate of 0.66 (Srivastava et al. 2014). The model was optimized with RMSprop, and training was terminated through early stopping with a patience of 5 or a maximum of 50 epochs. Early stopping used a validation split from the training data

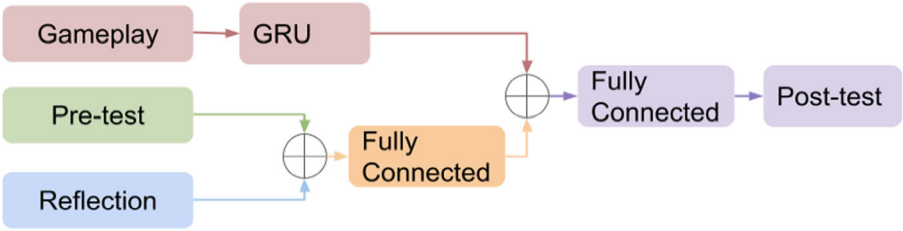

Fig. 5 Recurrent neural network architecture for predictive student modeling with student reflections

within the 10-fold cross-validation. The number of units in both fully connected layers was selected through hyperparameter tuning based on the validation data across 16, 32, 64, and 128 units. The number of units selected from hyperparameter tuning for the fully connected layer following the pre-test/reflection concatenation and the fully connected layer following GRU were 32/32 for the no-embedding model, 64/16 for the GloVe embedding model, and 128/128 for the ELMo embedding model. Finally, an ensemble was created using predictions from all seven models (e.g., naïve Bayes, KNN, linear SVM) using hard voting. The results for each model using either no-embeddings, GloVe embeddings, or ELMo embeddings are respectively presented in Tables 3, 4, and 5.

The accuracy of the predictions across each interval of a student's sequence from the best early prediction model, the ensemble of models with ELMo embeddings, is presented in Fig. 6. Generally, predictions look to stabilize after the first 20 min of gameplay with the exception of a set of students whose predictions continued to oscillate throughout the entirety of their sequence of predictions. There were several students whose predictions never reached an accurate prediction throughout the entire sequence. However, no discernible characteristics were found that distinguished this set of students from the larger set of students (e.g., no significant differences in pre-test performance or questionnaires).

Table 3 Early prediction models for classification on post-test without reflection data

Models	Early Prediction		Overall Classification				
	Standardized Convergence Point	Convergence Rate	F1 Score	Precision	Recall	Accuracy	AUC
Majority	.44	.55	**.71**	.55	**1.0**	.55	.50
Naive Bayes	.61	.61	.48	.58	.42	.59	.58
KNN	.40	.66	.62	.63	.62	.65	.65
LinearSVM	.44	**.68**	.65	**.69**	.61	**.70**	**.69**
LASSO	.45	.63	.62	.67	.58	.68	.67
Random Forest	.51	.56	.49	.55	.44	.58	.57
Adaboost	.52	.56	.52	.53	.51	.57	.56
Sequential NN	**.37**	.66	.58	.66	.52	.66	.62
Ensemble	.47	.62	.63	.67	.59	.68	.67

Table 4 Early prediction models for classification on post-test using GloVe embeddings for reflection data

Models	Early Prediction		Overall Classification				
	Standardized Convergence Point	Convergence Rate	F1 Score	Precision	Recall	Accuracy	AUC
Majority	.44	.55	**.71**	.55	**1.0**	.55	.50
Naive Bayes	.59	.60	.43	.60	.34	.59	.58
KNN	**.41**	.67	.61	.62	.60	.65	.64
LinearSVM	.46	.71	.61	.65	.57	.66	.65
LASSO	.43	.70	.62	.66	.58	.67	.66
Random Forest	.53	.61	.46	.55	.40	.57	.56
Adaboost	.57	.61	.48	.51	.45	.55	.54
Sequential NN	**.41**	.71	.67	.66	.68	**.69**	**.69**
Ensemble	.43	**.74**	.63	**.67**	.58	.68	.67

Discussion

Game-based learning environments show significant promise by providing adaptive learning experiences to individual students, but the effectiveness of adaptive learning experiences is dependent upon access to accurate, reliable student models. Predictive student modeling is critical for building effective adaptive learning environments, enabling the delivery of student-specific coaching and feedback. We examined several machine learning models for early prediction of student post-test scores within a game-based learning environment, CRYSTAL ISLAND. The predictive models utilized three alternate feature representations incorporating word embedding-based encodings of student reflection data: no reflection features, GloVe embeddings of reflection, and

Table 5 Early prediction models for classification on post-test using ELMo embeddings for reflection data

Models	Early Prediction		Overall Classification				
	Standardized Convergence Point	Convergence Rate	F1 Score	Precision	Recall	Accuracy	AUC
Majority	.44	.55	.71	.55	**1.0**	.55	.50
Naive Bayes	.56	.69	.58	.59	.57	.62	.62
KNN	**.36**	.72	.67	.66	.69	.70	.70
LinearSVM	.44	.73	**.73**	.71	.75	**.75**	**.75**
LASSO	.40	.73	.70	**.74**	.66	.74	.73
Random Forest	.45	.69	.60	.61	.59	.64	.64
Adaboost	.43	.69	.58	.63	.53	.65	.64
Sequential NN	.37	.74	.68	.68	.68	.71	.71
Ensemble	.41	**.77**	.71	**.74**	.69	**.75**	.74

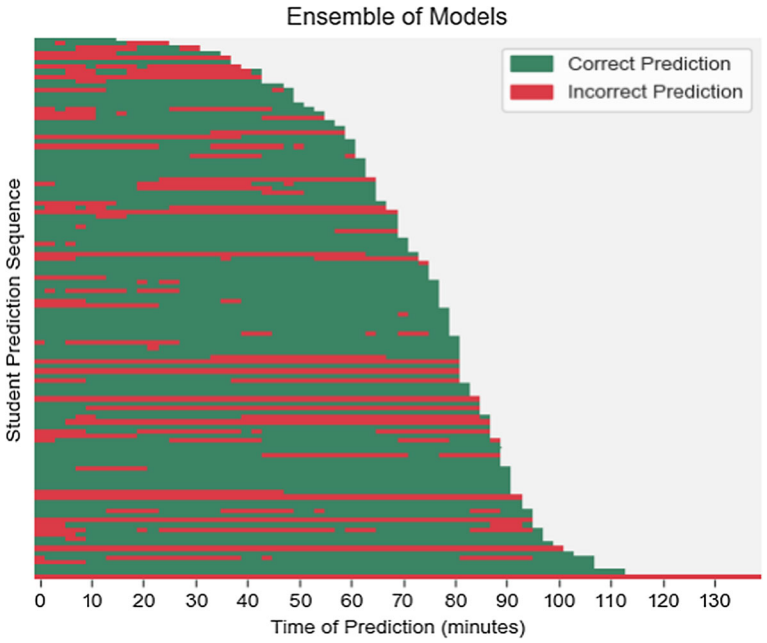

Fig. 6 Early prediction accuracy by student for the ensemble model using ELMo embeddings of reflection. Sorted by length for interpretability

ELMo embeddings of reflection. The linear SVM model and an ensemble consisting of seven models (e.g., naïve Bayes, recurrent neural network, linear SVM) that utilized ELMo embeddings of student reflection data demonstrated the best performance on overall classification accuracy and convergence rate, and substantially improved performance on F1 score and standardized convergence point relative to baseline models. The ensemble model achieved superior early prediction accuracy, while the linear SVM demonstrated superior overall accuracy.

The incorporation of features extracted from student reflections provided a substantial improvement to the model's predictive performance when utilizing ELMo embeddings. However, the GloVe embeddings provided little additional benefit compared to the no reflection feature representation. The large difference in classification accuracy across the GloVe and ELMo embeddings of reflection could be due to the additional capability of ELMo to encode contextual information into the feature representation. Through the use of contextual representations, ELMo is able to effectively disambiguate word sense, unlike GloVe, which forces a single representation for polysemous words. These results indicate that the incorporation of an additional information channel based upon student reflection text-based responses contributes toward the creation of predictive student models and student reflections can assist in inducing more effective predictive models. Additionally, the utilization of word embeddings allows for the integration of students' written reflection responses to reflection prompts without requiring resource intensive annotation by subject matter experts on the breadth and depth of the student's reflections.

The best performing models varied significantly across the six evaluation metrics. Standardized convergence point results revealed limited variability across the three

different feature representations, while convergence rate, F1 score, AUC, and accuracy all varied substantially. Given the use of a median split for post-test score, it was not surprising to see similar model selection across the accuracy and F1 score. Convergence rate consistently selected different models from the standardized convergence point, indicating a tradeoff between how early a model can get an accurate prediction and how reliably the early predictions are accurate. These results show the benefit of evaluating early prediction models using a diverse set of metrics, including convergence metrics borrowed from the research literature on planning and goal recognition (Min et al. 2020; Blaylock and Allen 2003; Min et al. 2016).

Predictive student modeling has great potential to satisfy the requirements of adaptive learning environments by providing efficient, accurate, and reliable student models. The creation of predictive models that leverage automatically generated representations of student reflections circumvents the need for resource-intensive annotations of students' natural language writing in the assessment process. The improved accuracy of early prediction models that utilize word embedding representations of student reflection allows for game-based learning environments that confidently adapt to student actions early in their problem-solving experience, thus providing greater opportunity to support the student. For example, if the predictive student model indicates that a student will perform poorly after 15 min, it may recognize that the student has missed several critical resources in the game, triggering a prompt to the student directing their attention toward the missed resources. Conversely, a confident prediction about student knowledge that occurs near the end of a student's gameplay session has limited utility, because there is limited opportunity remaining to intervene by providing adaptive coaching and support. Generally, early prediction models that can accurately and efficiently predict student learning outcomes offer the ability to alert instructors or drive adaptive scaffolding features to provide assistance to struggling students.

Limitations

A limitation of this study is that, in order to utilize the convergence metrics, students' post-test scores were converted into a binary variable. Ideally, post-test scores would be modeled directly as a regression problem, removing issues associated with classifying students near the median boundary as different classes. A second limitation is the treatment of students' reflection responses as an average of mean embeddings. This bag-of-words approach loses sequential information about how the student's reflections change over time during in-game problem solving and contextual information about word ordering in the written reflections. Investigating the use of convolutional or recurrent neural networks instead of average word embedding models is a promising approach to address this limitation, especially since word-order information is important when working with short responses. An additional limitation is the assumption that the content of prompted written reflection is representative of self-generated naturalistic reflection. Students with low motivation or engagement could produce responses to the written reflection prompts with limited authentic reflective thinking. Finally, while the incorporation of student written reflections using word embedding methods sidesteps the need for subject matter expert annotations of reflection quality, it does not provide a clear understanding of which features extracted from the written reflections lead to

improved accuracy of the predictive student model. On the one hand, it is possible that the models may encode key attributes of the students' reflections based on the word embeddings. On the other hand, it is also possible that the model simply utilizes surface-level attributes of the reflection texts, such as word complexity, word usage, and sentence length.

Conclusion

Predictive student models enable game-based learning environments to adapt to individual students' needs in real-time. Therefore, it is critical that the models be accurate and reliable. Reflection is a central component of self-regulated learning, and leveraging data on students' reflective thinking processes can greatly improve the accuracy of predictive student models. This study leveraged written reflections elicited throughout each student's interaction with a game-based learning environment by prompting the student to respond to embedded reflection prompts at different moments in the game. These written reflections were integrated into predictive student models, along with other data channels, by using pre-trained word embedding representations, such as GloVe and ELMo, to represent students' natural language responses to reflection prompts. Leveraging word embedding techniques removes the need for manual coding to extract relevant features from the written reflections. Additionally, the use of word embeddings that account for the context of each word in the student written reflections significantly improves the performance of predictive student models, as demonstrated using ELMo word embeddings. By combining student interaction data and written reflections represented using ELMo embeddings, predictive student models are able to accurately predict student learning outcomes at early points in the student's interaction with the game-based learning environment. Specifically, an ensemble of machine learning models employing ELMo word embeddings of written reflections demonstrated the best early prediction performance compared to other feature representations and model architectures. This approach grants the utilization of written reflections while eschewing the need to acquire resource-intensive annotations of the text. These findings highlight the importance of both capturing available data channels on student learning processes and encoding them automatically in ways that are beneficial for modeling student learning outcomes. Capturing these data channels not only enhances the promise of incorporating predictive student models into run-time game-based learning environments, but also provides insight into student learning processes.

Future Directions

There are several promising future directions for this work. Developing convergence metrics for continuous variables will be an important next step for evaluating early prediction in student models. Additionally, analyzing written responses can be extended beyond the direct application of unsupervised word embedding methods to more explicitly extracted qualitatively relevant reflection features. One approach is to pre-train word embeddings of written reflections to predict expert annotations of reflection characteristics (e.g., depth, categories, knowledge). The word embeddings from the pre-trained model could then be fine-tuned in a second supervised model predicting

student knowledge. Finally, it will be important to investigate the application of predictive student modeling within run-time game-based learning environments, setting the stage for empirically assessing the efficacy of early predictive models that integrate student reflection data to improve student learning outcomes.

Acknowledgements This research was supported by funding from the National Science Foundation (NSF) under Grant DRL-1661202. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Availability of Data and Material Data and materials created for this research are available upon request. Please direct all inquiries to the corresponding author.

Authors' Contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Michael Geden, Andrew Emerson, Dan Carpenter, and Jonathan Rowe. The first draft of the manuscript was written by Michael Geden, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This research was supported by funding from the National Science Foundation (NSF) under Grant DRL-1661202.

Compliance with Ethical Standards

Conflicts of Interest/Competing Interests No potential conflicts of interest.

Code Availability Code created for this research is available upon request. Please direct all inquiries to the corresponding author.

References

Acharya, A., & Sinha, D. (2014). Early prediction of students' performance using machine learning techniques. *International Journal of Computer Applications, 107*(1), 37–43.

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2017). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of the 5th International Conference on Learning Representations* (pp. 1–13).

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 *Proceedings of the 11th International Educational Data Mining*, (pp. 208-218).

Bannert, M. (2006). Effects of reflection prompts when learning with hypermedia. *Journal of Educational Computing Research, 35*(4), 359–375.

Bannert, M., & Reimann, P. (2012). Supporting self-regulated hypermedia learning through prompts. *Instructional Science, 40*(1), 193–211.

Baneres, D., Rodríguez-Gonzalez, M., & Serra, M. (2019). An early feedback prediction system for learners at-risk within a first-year higher education course. *IEEE Transactions on Learning Technologies, 12*(2), 249–263.

Baker, R., Corbett, A., & Aleven, V. (2008). More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (pp. 406–415).

Bell, A., Kelton, J., McDonagh, N., Mladenovic, R., & Morrison, K. (2011). A critical evaluation of the usefulness of a coding scheme to categorise levels of reflective thinking. *Assessment & Evaluation in Higher Education, 36*(7), 797–815.

Bisman, J. (2011). Engaged pedagogy: A study of the use of reflective journals in accounting education. *Assessment & Evaluation in Higher Education, 36*(3), 315–330.

Biswas, G., Baker, R. S., & Paquette, L. (2018). Data mining methods for assessing self-regulated learning. In D. H. Schunk & J. A. Greene (Eds.), Educational Psychology handbook series. Handbook of Self-regulation of learning and performance, 388–403. Routledge/Taylor & Francis Group.

Blaney, J., Filer, K., & Lyon, J. (2014). Assessing high impact practices using NVivo: An automated approach to analyzing student reflections for program improvement. *Research & Practice in Assessment, 9*, 97–100.

Blaylock, N. and Allen, J. 2003. Corpus-based, statistical goal recognition. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence* (pp. 1303–1308).

Brezovszky, B., McMullen, J., Veermans, K., Hannula-Sormunen, M. M., Rodríguez-Aflecht, G., Pongsakdi, N., Laakkonen, E., & Lehtinen, E. (2019). Effects of a mathematics game-based learning environment on primary school students' adaptive number knowledge. *Computers & Education, 128*, 63–74.

Boenink, A., Oderwald, A., De Jonge, P., Van Tilburg, W., & Smal, J. (2004). Assessing student reflection in medical practice. The development of an observer-rated instrument: Reliability, validity and initial experiences. *Medical Education, 38*(4), 368–377.

Botelho, A., Varatharaj, A., Patikorn, T., Doherty, D., Adjei, S. A., & Beck, J. E. (2019). Developing early detectors of student attrition and wheel spinning using deep learning. *IEEE Transactions on Learning Technologies, 12*(2), 158–170.

Boud, D., Keogh, R., & Walker, D. (2013). *Reflection: Turning experience into learning*. New York: Routledge.

Burrows, V., McNeill, B., Hubele, N., & Bellamy, L. (2001). Statistical evidence for enhanced learning of content through reflective journal writing. *Journal of Engineering Education, 90*(4), 661–667.

Chen, C., & Law, V. (2016). Scaffolding individual and collaborative game-based learning in learning performance and intrinsic motivation. *Computers in Human Behavior, 55*, 1201–1212.

Cheng, G. (2017). Towards an automatic classification system for supporting the development of critical reflective skills in L2 learning. *Australasian Journal of Educational Technology, 33*(4), 1–21.

Cisero, C. (2006). Does reflective journal writing improve course performance? *College Teaching, 54*(2), 231–236.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv*:1406.1078.

Costa, E., Fonseca, B., Santana, M., de Araújo, F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior, 73*, 247–256.

Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction, 22*, 9–38.

Emerson, A., Smith, A., Smith, C., Rodríguez, F., Min, W., Wiebe, E., Mott, B., Boyer, K., & Lester, J. (2019). Predicting early and often: Predictive student modeling for block-based programming environments. In *Proceedings of the 12th International Conference on Educational Data Mining* (pp. 39–48).

Falakmasir, M. H., Gonzalez-Brenes, J. P., Gordon, G. J., & DiCerbo, K. E. (2016). A data-driven approach for inferring student proficiency from game activity logs. In *Proceedings of the Third ACM Conference on Learning@ Scale* (pp. 341–349).

de Freitas, S. (2018). Are games effective learning tools? A review of educational games. *Educational Technology & Society, 21*(2), 74–84.

Gitinabard, N., Xu, Y., Heckman, S., Barnes, T., & Lynch, C. F. (2019). How widely can prediction models be generalized? Performance prediction in blended courses. *IEEE Transactions on Learning Technologies, 12*(2), 184–197.

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

Graesser, A. C. (2016). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education, 26*(1), 124–132.

Greene, J., Moos, D., & Azevedo, R. (2011). Self-regulation of learning with computer-based learning environments. In *New Directions for Teaching and Learning*, 107–115.

Greene, J., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research, 77*(3), 334–372.

Hume, A. (2009). Promoting higher levels of reflective writing in student journals. *Higher Education Research & Development, 28*(3), 247–260.

Jiménez, F., Paoletti, A., Sánchez, G., & Sciavicco, G. (2019). Predicting the risk of academic dropout with temporal multi-objective optimization. *IEEE Transactions on Learning Technologies, 12*(2), 225–236.

Jung, Y. Wise, A., & (2019). Teaching with analytics: Towards a situated model of instructional decision-making. *Journal of Learning Analytics, 6*(2), 53–69.

Karumbaiah, S., Baker, R., & Shute, V. (2018). Predicting quitting in students playing a learning game. In *Proceedings of the Twelfth International Conference on Educational Data Mining*, 167–176.

Kember, D. (1999). Determining the level of reflective thinking from students' written journals using a coding scheme based on the work of Mezirow. *International Journal of Lifelong Education*, *18*(1), 18–30.

Kember, D., McKay, J., Sinclair, K., & Wong, F. K. Y. (2008). A four-category scheme for coding and assessing the level of reflection in written work. *Assessment & Evaluation in Higher Education, 33*(4), 369–379.

Kim, Y., Almond, R., & Shute, V. (2016). Applying evidence-centered design for the development of game-based assessments in physics playground. *International Journal of Testing, 16*(2), 142–163.

Kinnebrew, J. S., Gauch, B. C., Segedy, J. R., & Biswas, G. (2015). Studying student use of self-regulated learning tools in an open-ended learning environment. In *Proceedings of the 20th international conference on artificial intelligence in education* (pp. 185–194). Cham: Springer.

Kostopoulos, G., Stamatis, K., & Kotsiantis, S. (2019). Multiview learning for early prognosis of academic performance: A case study. *IEEE Transactions on Learning Technologies, 12*(2), 212–224.

Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G. , & Dawson, S. (2018). Understand students' self-reflections through learning analytics. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 389–398).

Lai, G., & Calandra, B. (2010). Examining the effects of computer-based scaffolds on novice teachers' reflective journal writing. *Educational Technology Research and Development, 58*(4), 421–437.

Lester, J., Spires, H., Nietfeld, J., Minogue, J., Mott, B., & Lobene, E. (2014). Designing game-based learning environments for elementary science education: A narrative-centered learning perspective. *Information Sciences, 264*, 4–18.

Lykourentzou, I., Giannoukos, I., Mpardis, G., Nikolopoulos, V., & Loumos, V. (2009). Early and dynamic student achievement prediction in e-learning courses using neural networks. *Journal of the American Society for Information Science and Technology, 60*(2), 372–380.

Mao, Y., Zhi, R., Khoshnevisan, F., Price, T., Barnes, M., and Chi, M. (2019). One minute is enough: Early prediction of student success and event-level difficulty during novice programming tasks. In *Proceedings of the 12th International Conference on Educational Data Mining* (pp. 119–128).

Min, W., Baikadi, A., Mott, B., Rowe, J., Liu, B., Ha, E. Y., & Lester, J. (2016). A generalized multidimensional evaluation framework for player goal recognition. In *Proceedings of the 12th Artificial Intelligence and Interactive Digital Entertainment Conference* (pp. 197–203).

Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Smith, A., Wiebe, E., Boyer, K. E., & Lester, J. C. (2020). 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., and Lester, J. (2017a). Multimodal goal recognition in open-world digital games. In *Proceedings of the 13th Annual AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (pp. 80–86).

Min, W., Frankosky, M. H., Mott, B. W., Wiebe, E. N., Boyer, K. E., & Lester, J. C. (2017b). Inducing stealth assessors from game interaction data. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence in Education*, (pp. 212–223).

Meluso, A., Zheng, M., Spires, H. A., & Lester, J. (2012). Enhancing 5th graders' science content knowledge and self-efficacy through game-based learning. *Computers & Education, 59*(2), 497–504.

Minott, M. A., (2008). Valli's typology of reflection and the analysis of pre-service teachers' reflective journals. *Australian Journal of Teacher Education, 33*(5), 55–65.

Nietfeld, J., Shores, L., & Hoffmann, K. (2014). Self-regulation and gender within a game-based learning environment. *Journal of Educational Psychology, 106*(4), 1–13.

Olivé, D. M., Huynh, D. Q., Reynolds, M., Dougiamas, M., & Damyon, W. (2019). A quest for a one-size-fits-all neural network: Early prediction of students at risk in online courses. *IEEE Transactions on Learning Technologies, 12*(2), 171–183.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The achievement emotions questionnaire (AEQ). *Contemporary Educational Psychology, 36*(1), 36–48.

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543).

Pesare, E., Roselli, T., Corriero, N., & Rossano, V. (2016). Game-based learning and gamification to promote engagement and motivation in medical learning contexts. *Smart Learning Environments, 3*(5), 1–21.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2227–2237).

Piech, C., Spencer, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *Advances in Neural Information Processing Systems*, 505–513.

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

Poldner, E., Van der Schaaf, M., Simons, P. R.-J., Van Tartwijk, J., & Wijngaards, G. (2014). Assessing student teachers' reflective writing through quantitative content analysis. *European Journal of Teacher Education, 37*(3), 348–373.

Polyzou, A., & Karypis, G. (2019). Feature extraction for next-term prediction of poor student performance. *IEEE Transactions on Learning Technologies, 12*(2), 237–248.

Prilla, M., & Renner, B. (2014). Supporting collaborative reflection at work: A comparative case analysis. In *Proceedings of the 18th international conference on supporting group work* (pp. 182–193). New York: ACM Press.

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

Rogers, R. (2001). Reflection in higher education: A concept analysis. *Innovative Higher Education, 26*(1), 37–57.

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*(1–2), 115–133.

Sabourin, J., Mott, B., & Lester, J. (2013a). Discovering behavior patterns of self-regulated learners in an inquiry-based learning environment. In *Proceedings of the 16th international conference on artificial intelligence in education* (pp. 209–218). Berlin, Heidelberg: Springer.

Sabourin, J., Mott, B., & Lester, J. (2013b). Utilizing dynamic Bayes nets to improve early prediction models of self-regulated learning. In *Proceedings of the 21st International Conference on User Modeling, Adaptation, and Personalization* (pp. 228–241).

San Pedro, M., Baker, R., Gowda, S., & Heffernan, N. (2013). Towards an understanding of affect and knowledge from student interaction with an intelligent tutoring system. In *Proceedings of the 16th International Conference on Artificial Intelligence in Education*, 41–50.

Schön, D. A. (2017). The reflective practitioner: How professionals think in action. Routledge.

Schunk, D. H., & Greene, J. A. (2018). Historical, contemporary, and future perspectives on self-regulated learning and performance. *Handbook of Self-Regulation of Learning and Performance, 2*, 1–16.

Shen, D., Wang, G., Wang, W., Min, M., Su, Q., Zhang, Y., Li, C., Henao, R., and Carin, L. (2018). Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,* (pp. 440–450).

Shute, V. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction, 55*(2), 503–524.

Shute, V., 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.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.

Taub, M., Mudrick, N., Bradbury, A. E., & Azevedo, R. (in press). Self-regulation, self-explanation, and reflection in game-based learning. In J. Plass, B. Horner, & R. Mayer (Eds.), *Handbook of game-based learning*. Boston: MIT Press.

Thorpe, K. (2004). Reflective learning journals: From concept to practice. *Reflective Practice, 5*(3), 327–343.

Tsatsou, D., Vretos, N., & Daras, P. (2019). Adaptive game-based learning in multi-agent educational settings. *Journal of Computers in Education, 6*(2), 215–239.

Ullmann, T. (2011). An architecture for the automated detection of textual indicators of reflection. In *Proceedings of the 1st European Workshop on Awareness and Reflection in Learning Networks*, 138–151.

Ullmann, T. (2015). Automated detection of reflection in texts: A machine learning based approach. PhD Thesis.

Ullmann, T. (2019). Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education, 29*(2), 217–257.

Van den Boom, G., Paas, F., Van Merrienboer, J. J. G., & Van Gog, T. (2004). Reflection prompts and tutor feedback in a web-based learning environment: Effects on students' self-regulated learning competence. *Computers in Human Behavior, 20*, 551–567.

Van den Boom, G., Paas, F., & Van Merrienboer, J. J. G. (2007). Effects of elicited reflections combined with tutor or peer feedback on self-regulated learning and learning outcomes. *Learning and Instruction, 17*(5), 532–548.

Wang, L., Sy, A., Liu, L., & Piech, C. (2017). Learning to represent student knowledge on programming exercises using deep learning. In *Proceedings of the 10th International Conference on Educational Data Mining* (pp. 324–329).

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2016). Towards universal paraphrastic sentence embeddings. In *Proceedings of the 4th International Conference on Learning Representations* (pp. 1–19).

Winchell, M., Mozer, M., Lan, A., Grimaldi, P., & Pashler, H. (2018). Can textbook annotations serve as an early predictor of student learning? In *Proceedings of the 11th International Conference on Educational Data Mining* (pp. 431–437).

Winne, P., & Azevedo, R. (2014). Metacognition. In R.K. Sawyer (2nd Eds.), The Cambridge handbook of the learning sciences, (pp. 63–87). New York, NY.

Winne, P., & Hadwin, A. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Motivation and self-regulated learning: Theory, research, and applications* (pp. 297–314). Mahwah: Erlbaum.

Wong, F., Kember, D., Chung, L., & CertEd, L. (1995). Assessing the level of student reflection from reflective journals. *Journal of Advanced Nursing, 22*(1), 48–57.

Wu, M., Mosse, M., Goodman, N., & Piech, C. (2019). Zero shot learning for code education: Rubric sampling with deep learning inference. In *Proceedings of the 33rd International Conference of the Association for Advancement of Artificial Intelligence* (pp. 782–790).

Xiong, X., Zhao, S., van Inwegen, E., & Beck, J. (2016). Going deeper with deep knowledge tracing. In *Proceedings of the 9th International Educational Data Mining* (pp. 545–550).

Affiliations

Michael Geden[1] · Andrew Emerson[1] · Dan Carpenter[1] · Jonathan Rowe[1] · Roger Azevedo[2] · James Lester[1]

[1] Center for Educational Informatics, North Carolina State University, Raleigh, NC, USA

[2] Department of Learning Sciences and Education Research, University of Central Florida, Orlando, FL, USA