Received: 29 March 2022 | Accepted: 22 September 2022

DOI: 10.1111/bjet.13281

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Early prediction of student knowledge in game-based learning with distributed representations of assessment questions

Andrew Emerson[1] | **Wookhee Min**[1] | **Roger Azevedo**[2] | **James Lester**[1]

[1]Department of Computer Science, North Carolina State University, Raleigh, North Carolina, USA

[2]School of Modeling, Simulation, and Training, University of Central Florida, Orlando, Florida, USA

Correspondence

Andrew Emerson, AI Research Labs, Educational Testing Service, Princeton, New Jersey, USA.
Email: aemerson@ets.org

Funding information

Social Sciences and Humanities Research Council of Canada, Grant/Award Number: 895-2011-1006

Abstract

Game-based learning environments hold significant promise for facilitating learning experiences that are both effective and engaging. To support individualised learning and support proactive scaffolding when students are struggling, game-based learning environments should be able to accurately predict student knowledge at early points in students' gameplay. Student knowledge is traditionally assessed prior to and after each student interacts with the learning environment with conventional methods, such as multiple choice content knowledge assessments. While previous student modelling approaches have leveraged machine learning to automatically infer students' knowledge, there is limited work that incorporates the fine-grained content from each question in these types of tests into student models that predict student performance at early junctures in gameplay episodes. This work investigates a predictive student modelling approach that leverages the natural language text of the post-gameplay content knowledge questions and the text of the possible answer choices for early prediction of fine-grained individual student performance in game-based learning environments. With data from a study involving 66 undergraduate students from a large public university interacting with a game-based learning environment for microbiology, CRYSTAL ISLAND, we investigate the accuracy and early prediction capacity of student models that use

a combination of gameplay features extracted from student log files as well as distributed representations of post-test content assessment questions. The results demonstrate that by incorporating knowledge about assessment questions, early prediction models are able to outperform competing baselines that only use student game trace data with no question-related information. Furthermore, this approach achieves high generalisation, including predicting the performance of students on unseen questions.

KEYWORDS
game-based learning, natural language processing, predictive student modelling

Practitioner notes

What is already known about this topic

- A distinctive characteristic of game-based learning environments is their capacity to enable fine-grained student assessment.
- Adaptive game-based learning environments offer individualisation based on specific student needs and should be able to assess student competencies using early prediction models of those competencies.
- Word embedding approaches from the field of natural language processing show great promise in the ability to encode semantic information that can be leveraged by predictive student models.

What this paper adds

- Investigates word embeddings of assessment question content for reliable early prediction of student performance.
- Demonstrates the efficacy of distributed word embeddings of assessment questions when used by early prediction models compared to models that use either no assessment information or discrete representations of the questions.
- Demonstrates the efficacy and generalisability of word embeddings of assessment questions for predicting the performance of both new students on existing questions and existing students on new questions.

Implications for practice and/or policy

- Word embeddings of assessment questions can enhance early prediction models of student knowledge, which can drive adaptive feedback to students who interact with game-based learning environments.
- Practitioners should determine if new assessment questions will be developed for their game-based learning environment, and if so, consider using our student modelling framework that incorporates early prediction models pretrained with existing student responses to previous assessment questions and is generalisable to the new assessment questions by leveraging distributed word embedding techniques.
- Researchers should consider the most appropriate way to encode the assessment questions in ways that early prediction models are able to infer relationships

between the questions and gameplay behaviour to make accurate predictions of student competencies.

INTRODUCTION

Game-based learning environments offer engaging and effective experiences to enrich student knowledge (Mayer, 2019). Student knowledge can be assessed during game-based learning to inform proactive feedback (eg, hints) that can be given to students to further enhance their learning when they are struggling or wheel-spinning (Emerson et al., 2020; Sharma et al., 2019). To assess student knowledge, either in-game or post-test assessments are often administered to students to evaluate their content knowledge about the game-based learning environment's subject matter (eg, microbiology), at the time of assessment. Because assessing and modelling student knowledge during game-based learning can support real-time adaptations to the game, creating accurate student models that incorporate context from the desired subject matter is critical to employ adaptive support and to assess student learning.

Predictive student modelling is a set of techniques that aim to forecast student knowledge at early points in gameplay using interaction data generated by a student up to a specific moment (Geden et al., 2020). As a common approach to assess student knowledge in game-based learning, predictive student modelling induces student models automatically by using student behaviours extracted from their gameplay log files to predict learning outcomes. This eliminates the need for manual coding of student knowledge by domain experts, instructors or game designers. Gameplay log files have shown significant promise for the analysis of student behaviours in the context of assessment (Peters et al., 2021; Rowe et al., 2021). Recent years have seen a growing interest in predictive student modelling, with a particular focus on early prediction of student learning outcomes in game-based learning (Geden et al., 2021; Min et al., 2019). While previous work has aimed to predict student knowledge at early points in game-based learning, little work has explored the use of the content from the assessment questions in the student models.

The majority of studies investigating the assessment and prediction of student knowledge in game-based learning focus on one student population with a fixed set of assessment questions that are designed for that population (eg, multiple choice questions administered after the game) (Dever et al., 2021, 2022; Henderson et al., 2020; Taub et al., 2017, Taub, Sawyer, Lester, et al., 2020). This approach assumes that the content that will be assessed will not change. However, over time, teachers and test administrators may desire to refine the content of assessment questions to adhere to curriculum changes, prevent cheating due to the reuse of the same questions or due to changes in the problem-solving tasks in the game itself (Nietfeld, 2020). To adapt student models to new assessment questions and support more fine-grained inference about students' competencies, we argue for incorporating assessment question information directly into the predictive student models (ie, as features for machine learning predictive models). The field of natural language processing (NLP) has introduced methods (eg, word embeddings, distributed representation learning) for extracting salient information from textual data, such as assessment questions, to be used in machine learning. By incorporating this information into predictive student models that make inferences about learning outcomes, we argue that the combined value of the assessment questions and game logs will yield highly predictive student models. Additionally, this approach to predictive student modelling supports generalisation to new assessment questions for which a machine learning model has not been trained.

Related work

A wide range of investigations aim to assess and predict student knowledge in game-based learning (Plass et al., 2019). These approaches often leverage student gameplay behaviours to either forecast student knowledge at early points in gameplay or assess the knowledge that the student has learned during a full session of gameplay (Alonso-Fernández et al., 2020; Emerson et al., 2020; Min et al., 2019; Peters et al., 2021; Taub et al., 2017). In traditional intelligent tutoring systems, there are several methods for assessing student knowledge, including item response theory (Embretson & Reise, 2013), Bayesian knowledge tracing (Corbett & Anderson, 1994), deep knowledge tracing (Piech et al., 2015) and more recently, deep item response theory (Yeung, 2019). However, in game-based learning, assessing student knowledge often requires observing and modelling student interactions with the game. *Predictive student modelling* aims to predict future student competencies and learning outcomes using gameplay behaviour available up to the point of prediction (Geden et al., 2021). These models are similar to those of stealth assessment, which aims to infer student competencies and problem-solving skills using an evidence-centred assessment design framework (Kim et al., 2016; Shute et al., 2016; Shute & Rahimi, 2021). However, stealth assessment often requires the creation of probabilistic models with competency and evidence model variables, whereas predictive student models assess students without the need for preconstructed models.

To automatically analyse students' responses to assessments and to help generate and evaluate assessment questions that are relevant to the learning content, many studies have incorporated techniques from the field of NLP. Previous work has utilised NLP techniques such as *n*-gram analysis and part-of-speech tagging to better understand student learning outcomes (Sullivan & Keith, 2019) and to tailor learner support based on automated assessments (Moon et al., 2020). Using textual content from medical assessment questions, Xue et al. (2020) predicted the difficulty and response time of exam questions. Similarly, other studies have investigated advanced machine learning approaches that leverage the text of assessment questions to predict the difficulty of the questions (Benedetto et al., 2020; Huang et al., 2017). Other studies have leveraged student-generated text (eg, reflections) to better predict and understand student performance (Geden et al., 2021; Robinson et al., 2016; Su et al., 2018). However, these efforts do not utilise information from the assessment question text. Formative feedback can be enabled through the use of analysing student writing (Zhang et al., 2019). Feedback can also be enabled in the form of recommended reading, which can be automatically identified based on linking the student's knowledge deficiencies to the content of the learning material (Thaker et al., 2020). Formative assessment of written student reflections can provide deep insight into factors that influence a student's reflective thinking but often requires an expert-designed model to achieve this analysis (Liu et al., 2021). In sum, techniques from NLP can be leveraged to capture fine-grained detail about assessments, its relationship to student behaviours and student performance. While there has been limited research that aims to capture information about the questions themselves to predict student performance (eg, Condor et al., 2021; Huang et al., 2019), the current study aims to bridge the gap between NLP applied to assessment questions and predictive student modelling in game-based learning.

Research objectives and research questions

While there is a significant body of research that investigates predictive student modelling in game-based learning to predict student knowledge, limited work has leveraged the content

from the assessment questions to better model student knowledge at a fine-grained level. We address gaps in the literature by incorporating information extracted from the text of assessment questions to predict student knowledge. Specifically, we use word embedding techniques from the field of NLP to extract text from assessment questions that were administered after game-based learning. We then use the word embedding features as input to early prediction models that forecast student knowledge. By leveraging the combination of content from the assessment questions and game logs, we investigated the degree to which predictive student models were able to generalise the assessment questions for which the predictive student models were not trained. The current study reports on findings of an approach that uses *analytics for assessment* (Gašević et al., 2022). We investigate the following two research questions:

> **RQ1.** How well do early prediction models of student knowledge predict the performance of new students with a fixed set of assessment questions when incorporating game log data and content from the assessment questions?
>
> **RQ2.** How well do early prediction models of student knowledge predict the performance of the same students with a new set of assessment questions when incorporating game log data and content from the assessment questions?

We investigate how our predictive student modelling framework that utilises post-test questions as well as standard game interaction data can accurately infer students' performance on individual assessment questions compared to baseline approaches that utilise game interaction data only. We evaluate how this approach is generalisable with respect to both new student populations and new assessment questions.

METHODS

CRYSTAL ISLAND game-based learning environment

CRYSTAL ISLAND, the game-based learning environment used in this study, is designed for microbiology education (Figure 1). Students play the role of a medical researcher whose goal is to investigate a mysterious disease outbreak that is afflicting the research staff on a

FIGURE 1 CRYSTAL ISLAND game-based learning environment.

remote island. As part of the 3D interactive game environment, students can freely explore the game world and engage with the rich cast of characters and learning material. Student gameplay actions (eg, movement, dialogue and interaction with in-game objects) are all recorded along with the virtual location of where the action took place. Students collect relevant information and are tasked with solving the mystery by conversing with the game's characters; reading virtual books, papers and posters; and forming and testing hypotheses using the game's microbiology lab equipment. All in-game student actions, location and behaviour are automatically recorded in gameplay trace logs for subsequent analysis. Please see Taub, Sawyer, Smith, et al. (2020) for a detailed description of the learning environment.

Study participants and procedure

All predictive models in this work were induced using data captured during a data collection with undergraduate students interacting with the CRYSTAL ISLAND game-based learning environment in a laboratory setting. The study involved 66 participants (M = 20.0 years old, SD = 1.55), of which 23 (35%) were female. Before playing the game, students completed a 21-question multiple choice pretest that assessed their microbiology content knowledge (M = 11.59, SD = 2.72). The students played for an average of 67.9 minutes (min = 26.4 minutes, max = 159.8 minutes, SD = 22.0). After completing the game or for a maximum of 180 minutes, the students completed a microbiology assessment post-test (M = 14.35, SD = 2.86). The procedure is described in further detail in Taub, Sawyer, Smith, et al. (2020).

Multiple choice assessment questions

Traditional student models that predict future knowledge incorporate features of student prior knowledge (Min et al., 2019). In the approach presented in this paper, the predictive student modelling framework uses student performance on the pretest items as features and the items in the student's post-test assessment as target variables. We additionally extract the content of the individual post-test items as additional features to the models. If the predictive models can accurately predict student knowledge in a consistent manner after gameplay using data collected up to specific points (ie, high-fidelity early prediction models), then ideally no content knowledge assessments would be needed for future iterations of the game.

After playing the game, students completed a 21-item, four-option multiple choice post-test to capture acquired knowledge about microbiology content (M = 14.35, SD = 2.86, min = 8, max = 20). The post-test is the same assessment as the one administered prior to playing the game, but the questions are randomised to avoid practice effects. The validated post-test instrument was adopted from Nietfeld et al. (2014) and contains a mix of 12 factual and nine application questions. These questions were designed to assess microbiology knowledge for an undergraduate population. All information needed to answer the questions can be found in CRYSTAL ISLAND through books and papers, conversations with virtual characters and viewing informational microbiology posters. For example, one item is, 'What role do vaccines play in your immune system?' with the possible options of (A) 'They increase the number of lymph nodes', (B) 'They aid in the creation of new antibodies', (C) 'They prevent bacteria from growing' and (D) 'They prevent mutagens from activating', with the correct answer of B. One of the 21 questions had two correct answers, and the remaining had one correct answer. Student performance on each question item is displayed in Figure 2.

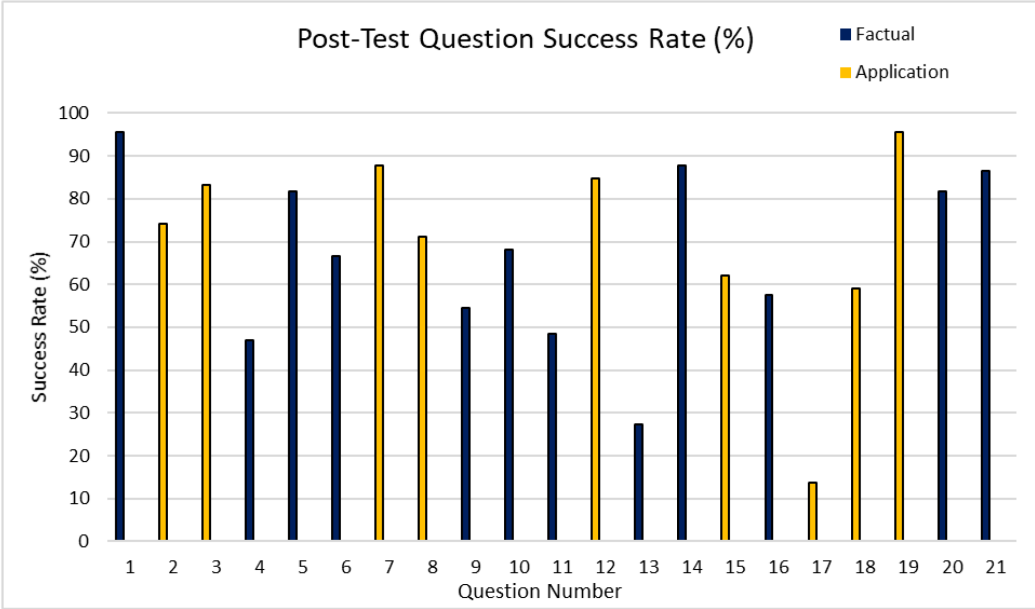

FIGURE 2 Student post-test performance by question number and type.

Data representation, coding and scoring

Assessment question representations

The post-test questions consist of two components: the question text and the text of the four answer options. To effectively represent this information for predictive student models, a series of preprocessing steps were conducted that are standard in natural language processing. First, all text was converted to lowercase and punctuation was removed. We then created three separate representations of the questions for further processing:

- *question-only (q)*: only uses the text of the question.
- *question-correct (q + c)*: concatenates the text of the question to the text of the correct answer.
- *question-all (q + all)*: concatenates the text of the question to all the text of both the correct and incorrect answers.

Each of these representations was then converted to word embedding-based encodings using one of two methods: 300-dimensional GloVe embeddings (Pennington et al., 2014) or 1024-dimensional ELMo embeddings (Peters et al., 2018). While GloVe embeddings offer a static representation for each word, ELMo provides a dynamic representation of a word considering the context of the sentence that includes the specific word. We evaluate if the contextual representations can improve the reliability of our early predictive student modelling framework. While these dimension sizes are quite different, we selected the largest pre-trained GloVe embedding available (300) and the smallest pre-trained ELMo embedding available (1024) to match the embedding sizes most closely. For each technique, we calculated the embedding for each word in the question configuration (ie, *q*, *q* + *c* and *q* + all) and averaged the resulting vectors to produce a single vector for the entire question. There are other newer word embedding techniques that are more computationally complex than GloVe

or ELMo being developed in the field of NLP. This study does not aim to evaluate which embedding approaches produce the most accurate models, but rather we aim to compare variants of contextual (eg, ELMo) and non-contextual (eg, GloVe) embeddings to show how word embeddings of assessment questions serve as a valuable input for predictive student modelling.

Student knowledge

For each pre-test and post-test question that were administered before and after CRYSTAL ISLAND, respectively, we used a binary representation to encode the correct (1) and incorrect (0) responses for each student. We include the pre-test responses as features to be included in the predictive model (k = 21), and we treat each individual post-test response as target variables to be predicted.

Gameplay data

As noted above, while interacting with CRYSTAL ISLAND, students' in-game actions and behaviours were recorded by logging software. To represent the gameplay interaction data, we categorised actions by their type, their location and any type-specific arguments that may accompany the action. For example, an action type includes reading a virtual book, the location could be the game's infirmary and the action arguments could be the title of the book that was being read. In total, there are nine action types, 24 locations and 97 possible action arguments that are further described in Geden et al. (2021). To represent these data for machine learning, we create a one-hot encoding of each action the student takes, the location of that action and possible arguments to that action. For example, when a student reads a book, a vector of size nine would be created, where all elements of that vector would be zero except for the position corresponding to reading a book. Subsequently, a similar vector would be created for the location of that action (k = 24) and a final vector for possible arguments (k = 97). These three vectors are concatenated together to create a single comprehensive vector that represents the student's action at that moment (k = 130). Over the course of the student's gameplay, a series of these vectors are created, all with the same size. To make early predictions of student performance, we aggregated the game action vectors by adding the vectors that occur up to the point of the prediction, making each element of the vector a sum of the actions and their corresponding arguments during that time period. We also include the student's current gameplay duration as a model feature, for a total of 131 gameplay features.

Early prediction models of student knowledge

To investigate the ability to which including content from assessment questions into predictive student models improves predictive accuracy, we constructed early prediction models of student post-test performance on individual assessment questions (Figure 3). We compared early prediction models that used gameplay behaviour logs only, which is the traditional data source for predictive student modelling work, to models that used gameplay logs together with a word embedding of the assessment question that is being predicted. When including the word embedding, we compared the performance of embedding the question only (q), question plus correct answer ($q + c$) and question plus all multiple choice options (q + all). We trained the early prediction models on data from the 66 students who interacted with CRYSTAL

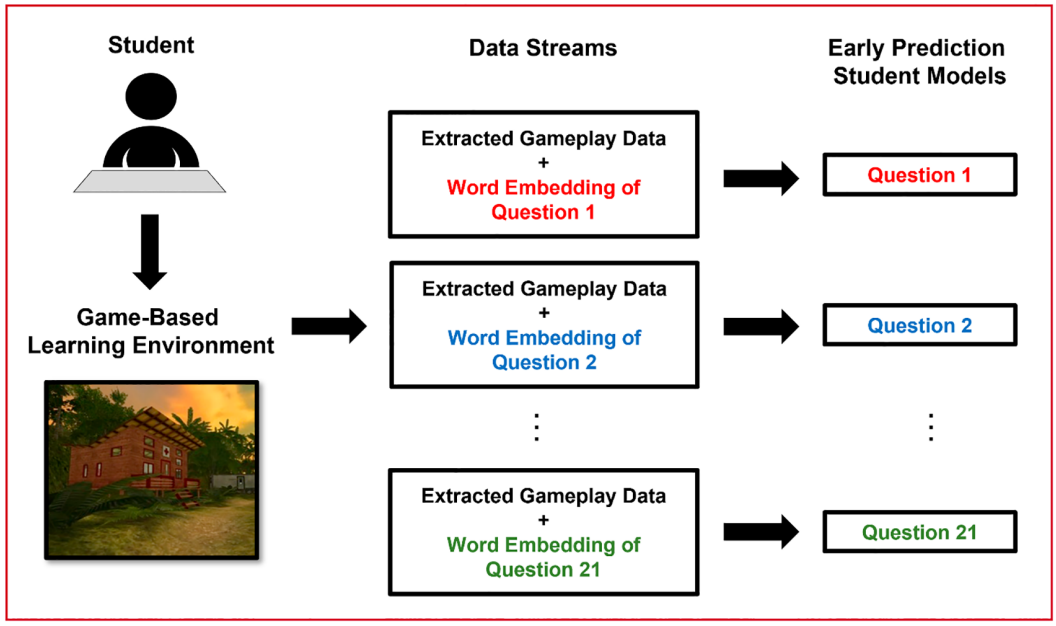

FIGURE 3 Overview of the predictive student modelling framework.

ISLAND. In total, there were 131 gameplay features and either 300 or 1024 word embedding features used for each predictive model. In addition, we incorporated 21 features derived from the pretest performance to each predictive model. The features from each source (ie, gameplay logs, assessment question content and pretest performance) are concatenated together for modelling.

To train and evaluate the early prediction models, we performed 10-fold cross-validation to enable the full set of 66 students and 21 assessment questions to be used for training and testing. For our first research question, we performed the cross-validation at the student level, that is, approximately one-tenth of the students were held out for evaluation, while nine-tenths of the students were used for training, during each fold. All assessment questions were used in each training fold when investigating this research question. For our second research question, we performed cross-validation at the assessment question level, that is, approximately one-tenth of the questions were held out for testing each fold. All students were used in each training fold for this research question. Our experiment designs evaluate our early predictive student modelling framework's generalisability with respect to students and questions by answering the first and second research questions respectively.

In this investigation, we encountered a common problem with having relatively little data. Specifically, the number of total possible features (k = 452 when using GloVe or k = 1176 when using ELMo) far outnumbered the number of students (n = 66) post-test questions (n = 21). To address this issue while still leveraging as much information from the full set of features as possible, we utilised principal component analysis (PCA; Abdi & Williams, 2010) as a way to condense the total set of features into a smaller set. PCA is a form of dimensionality reduction that transforms the original set of features into a reduced, orthogonal set. For the predictive tasks, we varied the PCA to be performed on the set of gameplay features and word embedding features either separately or jointly. This was conducted to determine if learned interactions between the assessment-based word embedding features and gameplay features assisted in predictive performance. We compared the performance of the predictive models when no PCA was applied (ie, all features available were utilised)

and with PCA final dimensions of 32, 64 or 128. For all predictive modelling in this manuscript, we utilised the random forest model for its flexibility and ability to perform well on a variety of classification tasks (Emerson et al., 2018). We vary the number of total decision trees used in the random forests and the minimum samples required for a leaf node as hyperparameters. In the Results section, the reported metrics are derived from the highest performing model configurations. The configuration includes the word embedding (ie, GloVe or ELMo), the type and dimensionality of PCA performed (ie, joint or separate; 32, 64 or 128) and the random forest hyperparameters. The machine learning models used in this work were written in Python 3 using the Scikit-learn library (Pedregosa et al., 2011).

Early prediction performance

To evaluate the early prediction performance of the student models, we used *standardised convergence point* (SCP) with the penalty of 1 (Min et al., 2016) and *convergence rate* (Blaylock & Allen, 2003). These metrics have been used previously for evaluating how well early prediction is conducted. SCP is a ratio calculated by averaging the number of observations for a model to predict in a consistently accurate manner divided by the sequence length for each sequence. Thus, it aims to convey how early the model can make accurate predictions consistently within a sequence of student actions. With the penalty parameter set to 1 (eg, Emerson et al., 2019; Henderson et al., 2021), SCP penalises sequences whose last instance predictions are incorrect (ie, non-converged sequences) by setting the SCP greater than 1, computed by (*the number actions* + 1)/*the number of actions*; for all converged sequences (eg, sequences whose last predictions are correct), the SCP falls within the range (0, 1]. As a result, smaller values indicate that the model is able to make accurate predictions earlier. Convergence rate indicates the percentage of how many sequences of student actions yield a final correct prediction. In combination, both of these metrics take into account the sequential nature of the predictions. Additionally, we evaluate the models using standard classification metrics: F1 score, accuracy and AUC. To predict student performance on the post-test questions, we predicted whether the students would correctly or incorrectly respond to the post-test question at every 2-minute interval during the game as well as at the end of the game. This time interval was chosen because it allows for students to perform a sufficient number of actions within the game for there to be a significant change in their gameplay feature representation at each interval, and it also enables a sufficient number of predictions per student. This time interval showed promise as a balanced value in prior work by Geden et al. (2021). This means that for the first time interval, the models will predict the student's performance on each of the assessment questions in the test set based on data available in the first 2 minutes of gameplay. The following intervals additionally include each subsequent 2 minutes of gameplay data.

RESULTS

RQ1. How well do early prediction models of student knowledge predict the performance of new students with a fixed set of assessment questions when incorporating game log data and content from the assessment questions?

To investigate how the addition of assessment question content affects the performance of early prediction models of student knowledge for new students, we designed predictive student models that made incremental predictions of individual post-test assessment

questions every 2 minutes during the student's gameplay. This approach aims to predict whether the student will answer each post-test question correctly or incorrectly at early points in the game. To make this prediction, we trained random forest binary classification models and compared the performance across different combinations of predictive features in terms of overall classification accuracy and early prediction convergence. In the comparisons, we tuned the word embedding used (ie, GloVe or ELMo), the PCA approach used and the random forest hyperparameters (ie, number of decision trees, minimum samples per leaf node). We report the 10-fold cross-validation results for each possible model, where the final results are the performance aggregated on the test folds. For this research question, the cross-validation was performed at the student level, meaning each training fold included all possible assessment questions and only a subset of the students, and the test folds included the remaining students. To investigate the extent to which adding assessment question content to the models, we compare this approach to two baselines: (1) A model that embeds assessment questions using a one-hot encoding called *One-Hot Questions* in Table 1 (ie, a vector with a '1' denoting the question and a '0' representing all other possible questions) and (2) a model that only uses student gameplay called *Gameplay-Only* in Table 1 (ie, no assessment question content used). These baselines aim to demonstrate that including assessment question content is effective, and specifically, representing assessment question content with word embedding models is effective when compared to modes that do not use this information. The results for the early prediction models for this research question are shown in Table 1. The * notation indicates that the model is statistically significantly better in performance for the specified metric in comparison to the *One-Hot Questions* baseline. Similarly, the ^ notation indicates the model statistically significantly outperforms the *Gameplay-Only* baseline. All statistical tests were conducted using the one-sided Wilcoxon signed-rank test between the performance of models in each cross-validation fold. This nonparametric test was chosen because the results from each fold are not normally distributed. Table 2 highlights the difference in early prediction performance when predicting either factual assessment questions or application questions using the best performing early prediction model (ie, lowest SCP).

TABLE 1 Post-test score early prediction results for the same questions but different students

Question representation	SCP[a]	CR[b]	F1	Accuracy	AUC
GloVe, Q	**0.320** *^	**0.727**	0.817	0.713	0.583
GloVe, Q + C	0.324*^	0.719	**0.820**	0.713	0.573
GloVe, Q + ALL	0.321*^	0.720	0.819	0.711	0.570
ELMo, Q	0.323*^	0.719	0.817	**0.725**	**0.623** *
ELMo, Q + C	0.324*^	0.726	0.817	0.721	0.609*
ELMo, Q + ALL	0.323*^	0.719	0.818	0.721	0.609*
One-Hot Questions	0.417	0.699	0.813	0.698	0.545
Gameplay-Only	0.467	0.701	0.793	0.699	0.614

[a]Standardised convergence point.

[b]Convergence rate.

*Indicates the model outperforms the one-hot questions baseline ($p < 0.05$);

^Indicates the model outperforms the gameplay-only baseline ($p < 0.05$).

Bolded values represent the highest performance for the given metric.

TABLE 2 RQ1 early prediction performance by question type using the GloVe, Q question representation

Question type	SCP[a]	CR[b]	F1	Accuracy	AUC
Factual	0.357	0.689	0.782	0.679	0.587
Application	**0.273**	**0.764**	**0.846**	**0.761**	**0.642**

[a]Standardised convergence point.

[b]Convergence rate.

Bolded values represent the highest performance for the given metric.

TABLE 3 Post-test score early prediction results for the same students but different questions

Question representation	SCP[a]	CR[b]	F1	Accuracy	AUC
GloVe, Q	0.411^	0.631	0.754^	0.632	0.522
GloVe, Q + C	0.404^	0.640	0.775^	0.653^	0.526
GloVe, Q + ALL	**0.395** ^	**0.657** ^	**0.781** ^	**0.659** ^	0.527
ELMo, Q	0.424^	0.627	0.770^	0.645	0.515
ELMo, Q + C	0.408^	0.651^	0.772^	0.645	0.512
ELMo, Q + ALL	0.421^	0.646^	0.768^	0.644	0.517
Gameplay-Only	0.486	0.592	0.701	0.597	**0.540**

[a]Standardised convergence point.

[b]Convergence rate.

^Indicates the model outperforms the gameplay-only baseline ($p < 0.05$).

Bolded values represent the highest performance for the given metric.

RQ2. How well do early prediction models of student knowledge predict the performance of the same students with a new set of assessment questions when incorporating game log data and content from the assessment questions?

To investigate how the addition of assessment question content impacts the performance of early prediction models of student knowledge for new assessment questions, we utilised the same processing and modelling pipeline described previously to compare the modelling conditions presented above. For this research question, the cross-validation was performed at the item level, meaning each training fold included all possible students and only a subset of the assessment questions, and the test folds included the remaining questions. We compare this approach to a baseline model that only uses student gameplay (ie, no assessment question content used). For this research question, we are unable to include a one-hot encoding baseline because it is not possible to represent individual unseen questions in the test sets using a fixed size vector. The results for the early prediction models for this research question are shown in Table 3. The ^ notation indicates the model is statistically significantly better in performance for the specified metric in comparison to the *Gameplay-Only* baseline. The statistical test used was again the one-sided Wilcoxon signed-rank test. Table 4 highlights the difference in early prediction performance when predicting either factual assessment questions or application questions using the best performing early prediction model.

DISCUSSION

In this paper, we demonstrated the ability to which adding assessment question content to early prediction models of student knowledge in game-based learning improved predic-

TABLE 4 RQ2 early prediction performance by question type using the GloVe, Q+ALL question representation

Question type	SCP[a]	CR[b]	F1	Accuracy	AUC
Factual	**0.393**	**0.662**	0.768	**0.654**	**0.551**
Application	0.397	0.648	**0.769**	0.651	0.534

[a]Standardised convergence point.

[b]Convergence rate.

Bolded values represent the highest performance for the given metric.

tive performance of those models. Similar studies have leveraged word embeddings of student-generated text in predicting performance (eg, Geden et al., 2021), but they do not incorporate the assessment question content. We compared early prediction models that leveraged the content of the question it was predicting to models that did not use this information. Specifically, we determined that by encoding the questions using word embeddings, early prediction models were better able to predict student performance on that question when compared to using a one-hot encoding of that question. Our results suggest that embedding assessment question content can improve early prediction accuracy for both new students and new assessment questions. This approach uses *analytics for assessment* (Gašević et al., 2022) and specifically aims to inform predictive student models in game-based learning to assess students using available gameplay log data more accurately.

For both research questions, we compared early prediction models that utilised word embeddings of assessment questions in addition to student gameplay data to models that used only gameplay data. Our results demonstrate that for both early prediction performance and overall classification accuracy, incorporating assessment question content improves overall predictive performance. Moreover, early prediction models incorporating distributed representations of assessment questions from either GloVe or ELMo embeddings outperform early prediction models that use one-hot encodings of the questions. While the one-hot encodings relay information about which question the model is predicting, the semantic content of the question is ignored. By leveraging distributed representation capabilities provided by word embeddings, the models are able to relate semantic information about the question (eg, concepts) that can be reused when predicting student success on multiple questions. This has significant implications for adaptive game-based learning environments that aim to tailor the game content to the needs of the student. For example, if a student is playing CRYSTAL ISLAND and the early prediction models forecast that the student will perform poorly on several questions related to viruses, the game can be adapted in real time to provide additional support related to this competency. Gameplay behaviours can also support this adaptation by indicating what the student has already tried. Using the virus example, consider the case where the student has read several books related to viruses and is still being predicted to perform poorly on the assessment questions related to viruses. The model may also detect that the student has not yet spoken with the virtual character who happens to be a virus expert, and in response, the game can encourage the student to speak with this character for further support. The adaptive support provided to the student can be applied to all question types (eg, procedural, conceptual, inferential) and can encourage the student to learn in different ways (eg, test a new hypothesis, read a new book, reflect on a particular challenge).

By leveraging the content of the questions in the models, the game is then better able to detect patterns in predicted student performance across subjects, types of questions (eg, factual or application) and question difficulty. We found that both GloVe and ELMo were effective in providing predictive value, with the GloVe model slightly outperforming ELMo in many cases. While ELMo does provide more contextual information about the question such as the word sense, GloVe has fewer features than ELMo (300 vs. 1024), which can

make it easier for the model to navigate signals in the data given the limited data set size investigated in this work. It is also possible that it is sufficient to encode the content of the question without accounting for the sequential nature of the words in the question, but this would require further investigation.

Early prediction models for new students with fixed questions

To answer RQ1, we constructed early prediction models of student performance on assessment questions by leveraging a combination of gameplay behaviour logs and the content of the assessment question for which the result is being predicted. Our results demonstrate that by incorporating distributed word embeddings of the question that is being predicted, we are able to increase the early prediction capacity (SCP and CR) and overall classification accuracy (F1, accuracy and AUC) compared to models that use no question content or one-hot encodings of the question. This indicates that when predicting the performance of new students on assessment questions that have existed in previous data, including information about the question is critical. However, we found that the early prediction models performed significantly better on application questions when compared to factual questions (Table 2). This finding could be due to several reasons. First, the 12 factual questions in this corpus have on average 9.67 words-per-question (24.83 including the answer options) compared to 15.22 words-per-question (33.56 including the answer options) for application questions. This indicates that application questions typically are longer, and it is possible that word embedding models are able to encode more contextual information from these additional words. Second, factual questions are typically in the form of 'true or false' (eg, 'Which of the following statements about genetic diseases is true?'), which makes it more challenging to capture semantic relationships with in-game content. Third, game-based learning environments feature an immersive, applied problem-solving experience where students are constantly applying knowledge they acquire. Game behaviour, in addition to question content, may be better suited to predict performance on application-style questions, especially when the questions have existed in previous student performance data. In practice, it is especially important to consider the types of questions that are being used to assess student performance when integrating adaptations driven by early prediction models. We did not find that including the answer options for the questions (ie, $q+c$, $q+$all) in addition to the question stem significantly improved predictive performance.

Early prediction models for fixed students with new questions

To answer RQ2, we utilised the same machine learning pipeline to predict the performance of existing students on assessment questions that had not been used in training. The results demonstrate that there is a consistent underperformance in terms of models' early prediction performance and predictive accuracy compared to the results we showed for RQ1. This phenomenon can be partially explained by the complexity of the predictive task being addressed with RQ2, where the early predictive student models do not have prior access to the content of the questions and must predict the performance of previously seen students on the new questions. However, our results showed that models that utilised assessment question content were able to significantly improve early prediction capacity and overall classification performance when compared to models that used gameplay alone. This has several implications for student modelling. There are very few studies that predict student performance on out-of-sample questions (eg, Condor et al., 2021; Huang et al., 2019). Typical machine learning models in educational settings will train classifiers on a fixed set of questions or competencies and predict the performance on those specific items.

Our approach enables educators, game designers and other educational stakeholders to continue to develop assessments and to create new, refined competencies on which to evaluate students. Adding these questions to the assessment would now no longer require a new study to be conducted, where students answer the new questions. The new questions could simply be encoded using the word embedding models, added as features to the model and the model would be able to generalise based on its training data consisting of other, perhaps related questions. For RQ2, we did not find that the performance differed significantly between types of questions, and this is somewhat intuitive. Without having used the exact question in training, it is more difficult for the model to discover relationships between that exact question type and the gameplay data, making the performance between question types less distinct. It is possible that with more questions and more student gameplay behaviours, a difference in performance by question type would be revealed. We did find that by including the answer options to the questions significantly improved predictive performance, which may allow the models to incorporate more context into making predictions.

Limitations

This study demonstrates the promise of leveraging assessment question content as predictive features to include in early prediction models of student performance on post-test questions. However, there are several limitations that should be noted. First, in encoding information about each assessment question, we did not attempt to relate questions to one another beyond the word embedding approaches for individual questions. As there are often several questions that evaluate the same competency, it would be potentially useful to leverage more information relating the questions to one another. A second limitation is that the word embedding approaches do not account for the difficulty of the question being predicted. Incorporating the difficulty of the question could help enhance the model's ability to capture distinct relationships between the question and gameplay behaviour (Huang et al., 2017). A final limitation is that our study only included 21 assessment questions. To be able to better understand the performance of early prediction models in this capacity, a larger sample size of questions would be ideal. A larger corpus of questions would create more variation in the word embeddings, which in turn would help the early prediction models capture more fine-grained differences in how students learn from game content and how this affects their performance.

CONCLUSION AND FUTURE WORK

Game-based learning environments offer significant opportunities for students to explore curricular content through multiple instructional approaches. These modes, such as reading virtual books, talking to characters and testing hypotheses, have distinct relationships with assessment questions, and predictive student models that leverage encoded information from the questions, in combination with student gameplay behaviours, achieve improved accuracies for predicting student performance. Moreover, predicting student performance at early points holds promise for driving adaptive scaffolding systems that can be built into game-based learning environments. We have introduced early prediction models that make use of student gameplay and assessment question content to predict student performance on assessment questions. Evaluation of models using this approach demonstrates that by leveraging assessment question content that has been encoded by word embeddings, early prediction models are able to outperform models that do not leverage this information, both when predicting the performance of new students on a fixed set of questions and also for a fixed set of students on new assessment questions.

Based on the findings of this study, there are several promising directions for future research. First, it will be important to evaluate this approach with a broader student population (eg, middle- and high-schoolers), set of assessment questions and game-based learning environments. Second, being able to induce relationships between questions when encoding aspects of the questions (eg, difficulty, type) will likely be helpful in predicting student performance on similar questions. This can be accomplished by investigating other techniques such as multitask learning from the field of natural language processing to relate similar information between questions (Bingel & Søgaard, 2017). A third direction for further study will be to investigate this approach on scenarios where both the students and the questions are different from the training set. In machine learning, this problem is often called zero-shot learning (Wu et al., 2019), where the model must be able to generalise to data with unseen labels. This case also has practical applications, where students from previous data collections interacted with the game and may have answered an old set of assessment questions. Then, the game designer and instructors refine the game and questions, and deploy the system to a new set of students. The same model may be used, but it now has to generalise further. A final direction for future work will be to investigate the types of adaptations and feedback the system will be able to deliver when it detects a student is struggling with a specific type of question or competency.

ACKNOWLEDGEMENTS

This study was supported by the funding from the Social Sciences and Humanities Research Council of Canada (SSHRC 895-2011-1006). Any opinions, findings and conclusions or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of the Social Sciences and Humanities Research Council of Canada.

CONFLICT OF INTEREST

There is no potential conflict of interest in this work.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ETHICS STATEMENT

The study data are not open due to human subject protection policies. This study was conducted with the IRB approval of North Carolina State University.

REFERENCES

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459.

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

Benedetto, L., Cappelli, A., Turrin, R., & Cremonesi, P. (2020). R2DE: A NLP approach to estimating IRT parameters of newly generated questions. In *Proceedings of the 10th International Conference on Learning Analytics & Knowledge* (pp. 412–421). Association for Computing Machinery.

Bingel, J., & Søgaard, A. (2017). Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics* (pp. 164–169). Association for Computational Linguistics.

Blaylock, N., & Allen, J. (2003). Corpus-based, statistical goal recognition. In *Proceedings of the 18th international joint conference on artificial intelligence* (pp. 1303–1308). Morgan Kaufmann Publishers Inc.

Condor, A., Litster, M., & Pardos, Z. (2021). Automatic short answer grading with SBERT on out-of-sample questions. In *Proceedings of the 14th international conference on educational data mining* (pp. 345–352). International Educational Data Mining Society.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278.

Dever, D. A., Amon, M. J., Vrzáková, H., Wiedbusch, M. D., Cloude, E. B., & Azevedo, R. (2022). Capturing sequences of learners' self-regulatory interactions with instructional material during game-based learning using auto-recurrence quantification analysis. *Frontiers in Psychology*, *13*, 813677. https://doi.org/10.3389/fpsyg.2022.813677

Dever, D. A., Wiedbusch, M., Cloude, E. B., Lester, J., & Azevedo, R. (2021). Scientific text comprehension during game-based learning: The impact of prior knowledge and emotions. *Discourse Processes*, *59*, 1–22.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

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.

Emerson, A., Sawyer, R., Azevedo, R., & Lester, J. (2018). Gaze-enhanced student modeling for game-based learning. In *Proceedings of the 26th Conference on User Modeling, Adaptation, and Personalization* (pp. 63–72). Association for Computational Machinery.

Emerson, A., Smith, A., Smith, C., Rodríguez, F., Min, W., Wiebe, E., Mott, B., Boyer, K. E., & 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). International Educational Data Mining Society.

Gašević, D., Greiff, S., & Shaffer, D. W. (2022). Towards strengthening links between learning analytics and assessment: Challenges and potentials of a promising new bond. *Computers in Human Behavior*, *134*, 107304.

Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., & Lester, J. (2021). Predictive student modeling in game-based learning environments with word embedding representations of reflection. *International Journal of Artificial Intelligence in Education*, *31*(1), 1–23.

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 34th AAAI Conference on Artificial Intelligence* (Vol. 34, No. 1, pp. 654–661). Association for the Advancement of Artificial Intelligence.

Henderson, N., Kumaran, V., Min, W., Mott, B., Wu, Z., Boulden, D., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E., & Lester, J. (2020). Enhancing student competency models for game-based learning with a hybrid stealth assessment framework. In *Proceedings of the 13th international conference on educational data mining* (pp. 92–103). International Educational Data Mining Society.

Henderson, N., Min, W., Emerson, A., Rowe, J., Lee, S., Minogue, J., & Lester, J. (2021). Early prediction of museum visitor engagement with multimodal adversarial domain adaptation. In *Proceedings of the 14th international conference on educational data mining* (pp. 93–104). International Educational Data Mining Society.

Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., Su, Y., & Hu, G. (2017). Question difficulty prediction for reading problems in standard tests. In *Proceedings of the 31st AAAI conference on artificial intelligence* (Vol. 31, pp. 1352–1359). Association for the Advancement of Artificial Intelligence.

Huang, Z., Yin, Y., Chen, E., Xiong, H., Su, Y., & Hu, G. (2019). Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, *33*(1), 100–115.

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

Liu, M., Kitto, K., & Shum, S. B. (2021). Combining factor analysis with writing analytics for the formative assessment of written reflection. *Computers in Human Behavior*, *120*, 106733.

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

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). Association for the Advancement of Artificial Intelligence.

Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Smith, A., Wiebe, E., Boyer, K. E., & Lester, J. C. (2019). DeepStealth: Game-based learning stealth assessment with deep neural networks. *IEEE Transactions on Learning Technologies*, *13*(2), 312–325.

Moon, J., Ke, F., & Sokolikj, Z. (2020). Automatic assessment of cognitive and emotional states in virtual reality-based flexibility training for four adolescents with autism. *British Journal of Educational Technology*, *51*(5), 1766–1784.

Nietfeld, J. L. (2020). Predicting transfer from a game-based learning environment. *Computers & Education*, *146*, 103780.

Nietfeld, J. L., Shores, L. R., & Hoffmann, K. F. (2014). Self-regulation and gender within a game-based learning environment. *Journal of Educational Psychology*, *106*(4), 961–973.

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

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543). Association for Computational Linguistics.

Peters, H., Kyngdon, A., & Stillwell, D. (2021). Construction and validation of a game-based intelligence assessment in Minecraft. *Computers in Human Behavior*, *119*, 106701.

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). Associations for Computational Linguistics.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In *Advances in neural information processing systems* (pp. 505–513). Curran Associates, Inc.

Plass, J., Mayer, R., & Homer, B. (Eds.). (2019). *Handbook of game-based learning*. The MIT Press.

Robinson, C., Yeomans, M., Reich, J., Hulleman, C., & Gehlbach, H. (2016). Forecasting student achievement in MOOCs with natural language processing. In *Proceedings of the 6th international conference on Learning Analytics & Knowledge* (pp. 383–387). Association for Computational Machinery.

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.

Sharma, K., Papamitsiou, Z., & Giannakos, M. (2019). Building pipelines for educational data using AI and multimodal analytics: A "grey-box" approach. *British Journal of Educational Technology*, *50*(6), 3004–3031.

Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, *116*, 106647.

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

Su, Y., Liu, Q., Liu, Q., Huang, Z., Yin, Y., Chen, E., Ding, C., Wei, S., & Hu, G. (2018). Exercise-enhanced sequential modeling for student performance prediction. In *Proceedings of the 32nd AAAI conference on artificial intelligence* (pp. 2435–2443). Association for the Advancement of Artificial Intelligence.

Sullivan, F. R., & Keith, P. K. (2019). Exploring the potential of natural language processing to support microgenetic analysis of collaborative learning discussions. *British Journal of Educational Technology*, *50*(6), 3047–3063.

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

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

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

Thaker, K., Zhang, L., He, D., & Brusilovsky, P. (2020). Recommending remedial readings using student knowledge state. In *Proceedings of the 13th International Conference on Educational Data Mining* (pp. 233–244). International Educational Data Mining Society.

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 AAAI Conference on Artificial Intelligence* (Vol. 33, No. 1, pp. 782-790). Association for the Advancement of Artificial Intelligence.

Xue, K., Yaneva, V., Runyon, C., & Baldwin, P. (2020). Predicting the difficulty and response time of multiple choice questions using transfer learning. In *Proceedings of the 15th workshop on innovative use of NLP for building educational applications* (pp. 193–197). Association for Computational Linguistics.

Yeung, C. K. (2019). Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory. *arXiv preprint arXiv*, 1904.11738.

Zhang, H., Magooda, A., Litman, D., Correnti, R., Wang, E., Matsmura, L. C., Howe, E., & Quintana, R. (2019). eRevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. In *Proceedings of the 31st AAAI Conference on Innovative Applications of Artificial Intelligence* (Vol. 33, pp. 9619–9625). Association for the Advancement of Artificial Intelligence.

How to cite this article: Emerson, A., Min, W., Azevedo, R., & Lester, J. (2022). Early prediction of student knowledge in game-based learning with distributed representations of assessment questions. *British Journal of Educational Technology*, *00*, 1–18. https://doi.org/10.1111/bjet.13281