What You Do Predicts How You Do

Prospectively Modeling Student Quiz Performance Using Activity Features in an Online Learning Environment

Emily Jensen
University of Colorado Boulder

Tetsumichi Umada
University of Colorado Boulder

Nicholas C. Hunkins
University of Colorado Boulder

Stephen Hutt
University of Pennsylvania

A. Corinne Huggins-Manley
University of Florida

Sidney K. D'Mello
University of Colorado Boulder

ABSTRACT

Students using online learning environments need to effectively self-regulate their learning. However, with an absence of teacher-provided structure, students often resort to less effective, passive learning strategies versus constructive ones. We consider the potential benefits of interventions that promote retrieval practice – retrieving learned content from memory – which is an effective strategy for learning and retention. The goal is to nudge students towards completing short, formative quizzes when they are likely to succeed on those assessments. Towards this goal, we developed a machine-learning model using data from 32,685 students who used an online mathematics platform over an entire school year to prospectively predict scores on three-item assessments (N = 210,020) from interaction patterns up to 9 minutes before the assessment as well as Item Response Theory (IRT) estimates of student ability and quiz difficulty. These models achieved a student-independent correlation of 0.55 between predicted and actual scores on the assessments and outperformed IRT-only predictions (r = 0.34). Model performance was largely independent of the length of the analyzed window preceding a quiz. We discuss potential for future applications of the models to trigger dynamic interventions that aim to encourage students to engage with formative assessments rather than more passive learning strategies.

CCS CONCEPTS

• **Applied computing** → Education; E-learning; • **Computing methodologies** → Machine learning.

KEYWORDS

Formative assessment, Item Response Theory, Machine Learning, Online Learning, Predicting Student Performance, Retrieval Practice

ACM Reference Format:
Emily Jensen, Tetsumichi Umada, Nicholas C. Hunkins, Stephen Hutt, A. Corinne Huggins-Manley, and Sidney K. D'Mello. 2021. What You Do Predicts How You Do: Prospectively Modeling Student Quiz Performance Using Activity Features in an Online Learning Environment. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21), April 12–16, 2021, Irvine, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3448139.3448151

LAK21, April 12–16, 2021, Irvine, CA, USA

ACM ISBN 978-1-4503-8935-8/21/04…$15.00
https://doi.org/10.1145/3448139.3448151

1 INTRODUCTION

Imagine the behavior of students who procrastinate studying for an exam scheduled for the next day. In a burst of late-night studying, they attempt to watch all of the online lecture videos that explain the concepts covered on the exam. The videos seem to make sense as they watch them. However, when they attempt a practice quiz, they realize that the lecture content went in one ear and out the other. After struggling through the quiz and getting a low score, they feel frustrated and anxious about their ability to perform well on their upcoming exam. Their motivation declines. If they can't figure out this little quiz, what is the point of putting in more time studying for what is going to be an even harder exam?

In a traditional classroom setting, teachers can structure how students engage with the course material, such as creating interactive rather than passive activities (see the Interactive, Constructive, Active, Passive (ICAP) framework in [17]). However, when left to self-regulate their learning, students often resort to fewer or less effective learning strategies such as passive viewing [18], verbatim notetaking or highlighting [30, 42, 51]. The problem is exacerbated in the context of online learning environments, such as Massive Open Online Courses (MOOCs) and Khan Academy, which are increasingly common, especially in the age of a global pandemic. In these platforms, students are often provided with structured content in the form of ordered lessons. However, student interaction time and learning strategies are largely unregulated. Students must allocate their time between the various learning activities such as watching videos and taking assessments. Although viewing well-designed instructional videos can offer feelings of engagement, and competence [2, 50], in reality, very little actual learning occurs, especially for difficult conceptual content [36].

In contrast, research has demonstrated the benefits of retrieval practice on long-term retention (the so-called testing effect) and end-of-course outcomes [3, 29, 38, 54, 55]. During retrieval practice, students practice retrieving information from memory, often in the form of a multiple-choice quiz, flashcards, or free-response prompts. Compared to passive strategies such as re-reading, retrieval practice forces students to evaluate weak points in their knowledge when they cannot recall the correct answer (which allows them to then focus on these areas when studying) and promotes deeper memory encoding [40]. Unfortunately, when given a choice, students prefer

more passive strategies such as re-reading or taking notes, instead of engaging in the more effortful retrieval practice [30, 42]. Even students aware of the benefits of retrieval practice feel more confident in their knowledge after passive activities such as re-reading [55].

It is clear that students need to develop skills to effectively regulate their learning [27, 50, 63, 67]. One approach is to help students develop self-regulatory skills to become more effective learners. For example, the MetaTutor project [7, 8] proposes a framework to detect, trace, model, and foster self-regulated learning habits. This work uses an animated tutor as an external regulatory agent, providing feedback and scaffolding to help the student help themselves. Similar frameworks have been suggested for other computerized learning environments such as Bettys Brain [52] and for blended learning environments [5].

An alternative approach, which we explore here, is to provide nudges [60] or suggestions for students about how they should be using their time. Specifically, we seek to take an initial step towards encouraging students to engage in retrieval practice in the form of short, formative assessments. These nudges should be delivered at appropriate times so that they are not disruptive or harmful. Nudging too frequently would be annoying and disruptive, whereas nudging at inappropriate times might even be detrimental. In particular, it is not likely to be beneficial to suggest students engage with assessments for which they have little likelihood of success, and this can negatively impact engagement and discourage students from learning [3]. An ill-timed assessment might also reduce motivation to engage in future assessments.

But what if we can predict when a student is likely to succeed on an assessment? In this case, a well-timed nudge to successfully complete an assessment can provide positive feedback, which should positively increase motivation to use these assessments in the future [37]. In addition to the aforementioned cognitive benefits on learning [59, 62, 64], research suggests that this form of interpolated testing can improve engagement by reducing attentional lapses and improving motivation [58]. This is not to propose a gate-keeping mechanism for taking a quiz based on an anticipated low score; there are a variety of factors influencing quiz performance and students should be allowed to attempt a quiz when it aligns with their goals and motivations.

The purpose of this paper is to explore the possibility of prospectively predicting student success on a short, formative assessment as an initial but critical step in implementing intelligent, well-timed nudges. Information on potential student success on an assessment can be obtained from a variety of sources. Previous work uses student modeling approaches, including Knowledge Tracing [9, 66], Performance Factor Analysis [48], and Item-Response Theory (IRT) [16], which identifies student mastery of specific concepts or skills based on their previous responses to problems/items. As we elaborate below, this approach requires frequent assessments to inform the underlying learner models, and such data might not be available when such assessments can be infrequent or even non-existent as in the case of unstructured online learning environments. They also require parameterized models of domain knowledge and assessment items, which might not be available in MOOCs and other online learning platforms.

Alternatively, we explore whether student actions immediately preceding an assessment can predict their subsequent performance. Based on the ICAP framework and retrieval practice (see above), we expect that participation in more active and constructive activities (such as taking quizzes) is associated with better performance on following assessments than more passive strategies. This is in line with [4], where the authors found that completing assessments and using expert-labeled active participation were most associated with end-of-course success. Additionally, [31] found students who completed relatively more interactive activities performed better at the end of a course than students that watched more videos or read more pages. Accordingly, we investigate whether student performance on low-stakes assessments in an online learning platform can be prospectively predicted based on their prior interaction patterns with the platform. We also examine if accuracy can be improved by combining interaction data with model-based estimates (i.e., IRT).

We trained our models using a large, diverse data set of real-world interaction data from students (N = 32,685) of varying demographics, knowledge backgrounds, and classroom experiences who engaged in an online math learning platform for an entire school year. If we can successfully predict student performance on a quiz based on preceding actions, we can then develop real-time nudges to balance student motivation, engagement, and learning.

1.1 Related Work

To keep scope manageable, we focus on studies conducted in online or hybrid learning environments. Previous research on predicting student performance has focused on outcomes at a variety of levels. Most broadly, some researchers have attempted to predict drop-out from MOOCs [20, 31] or a summative measure of course performance such as course grade or final exam score [4, 6, 24, 35, 53, 57] (for a review, see [39]). Our current work is not focused at this level of prediction, instead aiming to provide more specific, actionable feedback throughout a course.

In contrast, other lines of research have predicted student performance on individual assessment items [12, 15, 43–46, 56], programming problems [25, 33], or entire assessments [21, 24]. For example, in [24], the authors predicted scores on intermediate assessments using personalized linear regression models [23] and detailed features from interaction logs. Similar to [53], they found that viewing course materials, in addition to cumulative GPA and current course grade, was the most important indicator of student success. As discussed above, we choose to focus our predictions on intermediate scores such as these (compared to a whole-course level) so that we can give meaningful feedback to students at timely intervals over an entire course.

Previous approaches have used a variety of features and methods to predict student success. One active line of research uses Bayesian Knowledge Tracing [19] and related methods [21] to model student mastery of specific concepts and skills based on performance on individual problems [9, 43–46, 66]. Unfortunately, this strategy is not easily applicable outside of a problem-solving environment where student performance on problems and other assessment items are sparse. In addition, this approach requires domain experts to develop knowledge models specific to the platform or developing

a latent knowledge model [32]; both options might not be available for online learning platforms.

Other research has predicted student performance using features external to the immediate learning environment. These include student grades [24], prior performance within the course [15, 25], historical course and instructor information [24], and student demographics [24]. For example, perhaps unsurprisingly, [25] found one of the strongest positive predictors of success on programming problems was prior performance. Additionally, in [15] the authors predicted student performance on individual assessment items by incorporating an estimated measure of student current knowledge. While these external factors may influence student performance, they do not take into account the student's learning behavior, which we aim to target with timely feedback. Further, using certain features such as demographics can encode systematic biases that may impact model predictions in a manner that affects marginalized groups [22, 68].

Finally, a growing body of work has focused on predicting student performance using information from the immediate learning environment. Some work has found success using simple counts of activities on the learning platform [4, 31, 53, 57] or engineered interaction patterns and sequences [4, 12, 15, 24, 25, 33]. For example, in [53] the authors used stepwise multiple regression to predict total test scores in a course using activity frequencies. They found the main predictor of success was total page hits, which indicates the overall level of interaction with course content. In [31], the authors used a causal inference system to analyze course dropout and end-of-course scores using activity frequencies. They found that students that completed relatively more interactive activities performed better than students that watched more videos or read more pages. While these activity features are similar to those used in this work, we choose to focus on context-specific actions leading up to a quiz rather than an aggregate over an entire term. Additionally, [12] extracted generic video-watching behaviors to predict whether students would answer correctly on the first attempt of a problem related to the video. They found that interactions with the video (e.g., pausing, rewinding) increases the chance of success on the problem. However, this work only considers the actions in the context of specific video-problem pairings; we choose to examine a broader context of student activity leading up to the start of an assessment.

Approaches using data from the student's immediate learning environment are particularly promising as they can be developed to be system-agnostic and used in different contexts. For example, [4] analyzed end-of-course scores using three different system-independent interaction patterns: between agents (students, teachers, content), frequency of resource use, and active versus passive interaction. Most importantly, they found that participating in assessments and using active participation were most associated with end-of-course success. More generally, [57] was able to predict grade sequences using unusual spikes in activity during a course.

Two studies in particular are most similar to the current work. First, [15] predicted performance on items using measures of prior performance on specific concepts, current knowledge, item difficulty level, and engineered activity features. While this method combines both information about the current learning session as well as external factors to predict student performance, the features used rely on expert-constructed learning units and difficulty ratings of easy or difficult. In this work, our models generally approximate student ability through their past performance on prior quizzes rather than tracking their mastery of individual concepts. Our results were obtained with a psychometric IRT model of item difficulty and estimated student ability. Additionally, [25] used general activity features to predict student completion of programming problems. While the features used (such as interacting with the platform) are similar to those used here, the authors focused on predicting student success *after* the start of the problem. In this work, we aim to prospectively predict student success by considering features before the start of an assessment.

1.2 Contribution & Research Questions

Our overarching goal is to develop models of student performance on embedded quizzes to deliver timely nudges to encourage effective learning strategies. This paper takes an initial step in this direction by providing a proof-of-concept that we can model student performance from their immediate learning context. To do this, we use log data to prospectively predict performance on 210,020 quiz attempts from 32,685 students. We go beyond previous work by combining general activity features leading up to the start of a quiz with IRT factors such as estimated student ability and quiz difficulty that do not rely on domain expert knowledge. Additionally, we consider models using different combinations of these features. If successful, this approach can be used in an intervention that selectively encourages students to attempt assessments when they are likely to be successful. We address five specific research questions (RQ).

RQ1. To what extent can we prospectively predict student success on a short assessment (quiz) using only information from the immediate learning context? Work such as [25, 33] has predicted student success using activity after students have begun to solve an individual item. Instead, we aim to understand a student's ability to succeed before they even start the assessment. To answer this question, we build models using activity features collected from the 3-minute window (we experimented with other window lengths) leading up to the start of a quiz. Our approach does not require a pre-determined model for each individual student, but relies on activity patterns alone, as these are more easily available in online learning platforms.

RQ2. Within the activity patterns used in RQ1, what is the influence of prior retrieval practice on subsequent quiz performance? To investigate this question, we trained separate models that either considered only quiz-related activity or non-quiz activity. We found that models using only quiz-related activity from previous quizzes predict student performance better than those using non-quiz activity (e.g., video viewing), but there is information content in both types of features.

RQ3. How accurately can we model student performance using information independent of the immediate learning session? We answer this question by comparing the activity patterns with IRT measures of student ability and quiz difficulty. We achieved the best predictive performance by combining the two approaches.

RQ4. How much context is needed to predict student performance? We comparatively evaluated longer context windows leading up to the start of a quiz but found no difference in predictive performance as we consider actions farther from the start of a quiz.

RQ5. What features are most influential in predicting student performance? We analyze the relative importance of the individual features from the machine-learned models as well as more interpretable coefficients from a linear regression. Proxies for quiz difficulty and student ability were among the most important in addition to activities related to previous retrieval practice and the discussion board.

2 METHOD

The data used in this study was collected through the Algebra Nation platform, described below. Previous work [26, 28, 41] has investigated interaction patterns and their effects on students using the platform, but these use data collected over a different school year and for a different modelling problem than is considered here.

2.1 Algebra Nation

Algebra Nation is a large-scale online math learning platform developed by Study Edge, an educational software and tutoring company. Students can access Algebra Nation through the website (https://www.algebranation.com/) or a mobile app. Over 150,000 students use Algebra Nation each semester in Algebra 1, Geometry, and Algebra 2. Each domain uses the same interaction framework and range of activities (discussed below). Content for each domain is aligned with Florida state education standards. In this paper, we focus on data from Algebra 1, which is organized into 10 sections which each cover 6 to 14 specific topics. A suggested topic sequence is provided to students; however, students are free to interact with Algebra Nation in whatever way they choose or based on what their teachers require.

Algebra Nation provides short *video* lectures for each topic within a section. Each lesson is recorded by multiple human tutors, where each tutor provides a unique perspective on the topic along with different levels of expressiveness and technical detail. Students can choose any of the tutors and can switch tutors at any time.

The Algebra Nation community interacts through a *Discussion Wall* (separately for each math subject). Students post requests for help, which may be answered by other students or study experts (teachers working with Algebra Nation). Students who provide helpful guidance are awarded *karma points* by study experts. Karma rankings are shown in a leaderboard and monthly prizes are given to students with the highest karma.

Algebra Nation also offers a *Test Yourself! Practice Tool* (TYS), which delivers a randomly selected set of 10 items on the selected section and covers several video topics. These are selected from a pool of items aligned with state standards and reviewed by content experts and teachers for content validity evidence. For internal structure validity evidence, the 2PL IRT model fit well to the item data, and multiple psychometric methods were used to obtain unbiased item parameters for the state level population relevant to our data [65]. Additionally, multiple psychometric analysis internal to our research project found evidence for criterion-related validity for the Test Yourself items. As one example, responses on Test Yourself items from a previous year of statewide data were correlated to scores on the end-of-year state standardized Algebra test. Out of 531 items, 526 showed positive correlations, with an overall mean point biserial correlation of .34 and standard deviation of 0.12. Students answer items through open text boxes or multiple-choice options, which are then automatically graded (Figure 1a). Students may review their performance on the items, view solution videos, and revisit items and topic videos (Figure 1b).

Algebra Nation recently introduced additional *Check Your Understanding* (CYU) quizzes. These are short quizzes (three fixed items each) on a specific video topic. Items are specific to the content covered in each video and do not overlap with the Test Yourself item pool. These quizzes were designed for students to practice what they just learned in the videos and to obtain formative feedback to help guide their learning session. It is distinct from the more elaborated Test Yourself practice assessment, which covers an entire section rather than a specific topic. In this work, we focus on predicting performance on these CYU quizzes. The CYU item content was developed and examined by content experts in the same manner as Test Yourself items, and the 2PL fit well to the data of all but five items, which were removed from the analysis. In addition, the IRT ability scores from each of the CYU tests in Algebra Nation in the Spring of 2020 were positively correlated with prior year state standardized mathematics scores (noting that Algebra standardized test scores were unavailable for Spring 2020 due to COVID-19 testing cancellations).

Each quiz contains three items, which are often related to real-world situations and are presented in the form of fill-in-the-blank, interpreting a graph, or describing why a statement is true. Although each quiz contains three items, each item can contain multiple sub-items which may rely on the answers of the previous parts. Quizzes are graded out of three, one point per item; if a student answers one sub-part of a item incorrectly, then the entire item is marked as incorrect by the platform. Students are free to move between items during the quiz. When they submit their final responses, they are shown their overall score as well as an option to view feedback on each item. Students can then review the correct answer and solution steps for each item, re-watch the topic video explaining the concepts of the quiz, or attempt the quiz again.

2.2 Activity Features

We used features that did not rely on domain-specific content (e.g., watching a video lecture on factoring polynomials), quiz items (e.g., solving a system of equations), or specific student-generated text (e.g., a request for help on the discussion wall). Our activity features represent counts of 22 actions aggregated across the 3-minute window immediately leading up to the start of a quiz (we vary the window size later). Figure 2 illustrates the instance-building procedure. The activity features can be divided into four subcategories: watching videos, taking assessments, interacting with the discussion wall, and Algebra Nation onboarding. The assessment features might include both previous Check Your Understanding or Test Yourself assessments as part of retrieval practice. Recall that all features are computed prior to each quiz, so they reflect actions on earlier assessments. However, we only consider a student's first

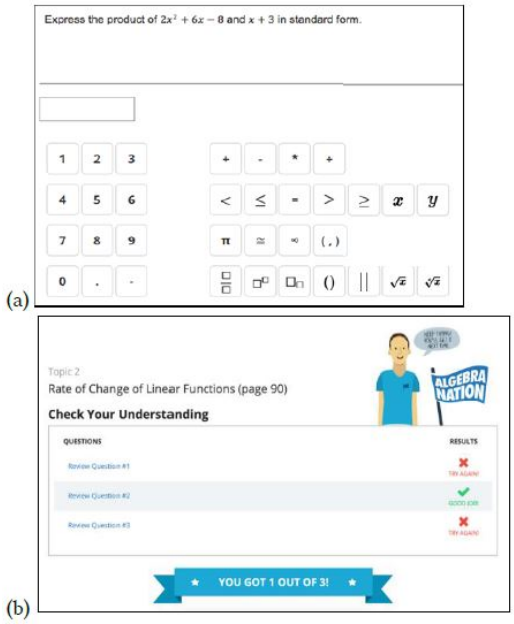

Figure 1: (a) Open-ended Check Your Understanding item (b) Feedback after submitting a Check Your Understanding quiz.

attempt of each quiz, thereby excluding cases where students immediately retake a quiz for a higher grade, which is particularly easy to do as the correct answers were available following the first attempt.

Table 1 contains a full list of activity features and average occurrences within each window. The first set of activities relates to watching lecture videos. Specifically, once a student starts playing a video, they can navigate within the video (pause, seek, resume), turn on captions, and exit the video environment. Video lengths vary depending on the selected tutor; videos providing a shorter review range in length from 8 to 20 minutes, while videos giving an in-depth discussion of the topic range in length from 16 to 29 minutes. The next set of activities relates to taking assessments. Students can answer or return to individual items, finish an assessment, and load or unload the general assessment environment. Once the results of the assessment are displayed, students have the option to review the answers to individual items (whether they were correct or incorrect) as well as watch videos for item solutions and the general topic of the assessment. Students can also open the discussion wall, load additional wall posts, search for previous posts, or post their own item or response. If a student posts a helpful response, they are awarded *karma points*. Finally, students can watch biographical videos about the expert tutors.

In general, these features are very sparse, where students only complete a few activities within the 3-minute window. The most frequent activities on average were answering a quiz item or reviewing an incorrect answer on a quiz. For some actions, particularly video viewing, the database sometimes recorded too many actions within the 3-minute window (e.g., pausing a video hundreds of times). Although these outliers were rare, we accounted for this by clamping each feature to a 20-count maximum per one-minute interval similar to the procedure in [26].

2.3 IRT Features (Student Ability and Quiz Difficulty)

In addition to the action counts, which represent the immediate learning session, we also computed student ability and overall quiz difficulty using an IRT framework. The two-parameter logistic model (2PL [10]) was chosen due to the binary nature of the Algebra Nation item responses and the relative parsimony of the model compared to many other IRT models for binary data. The 2PL allowed us to evaluate and utilize two core measurement features of items, difficulty and discrimination, without introducing additional parameters that can lead to both technical and interpretational challenges, as happens, for example, when freely estimating a lower asymptote parameter [34, 47]. In addition, our data mining approach to locating unbiased item parameters was tractable within the 2PL framework. Details on how difficulty and discrimination were calculated can be found in [65].

Using this framework, section-specific student ability (θ) was estimated based on the most recent quiz performance in the section, quiz item difficulty (β), and quiz item discrimination (α). We use fixed-parameter calibration, as commonly done in computer adaptive testing [61], by fixing the item parameters to the values located in [65] while estimating a person-level latent trait parameter with expected-a-posterior estimation (EAP). We used EAP due to its speed in estimating unidimensional traits due to the non-iterative nature of the estimation procedure [13]. The estimation is

Quiz ID	Play	Pause	Rewind	Forward	Quiz	Post
1	3	2	1	0	0	1

Figure 2: Example of aggregating action counts in a 3-minute window preceding a quiz.

Table 1: Distribution of activity features for 3-minute windows leading up to starting a quiz. (N = 210,190). Feature importance for the Random Forest and Linear Regression combined models are also reported after averaging across 5 cross-validation folds (standard deviations across folds were near zero).

Activity Description	Mean	SD	Median	Random Forest Importance	Linear Regression Coefficient
Start watching video	0.10	0.45	0	0.00	-0.01
Pause video	0.30	1.37	0	0.00	0.02
Resume a paused video	0.27	0.35	0	0.00	-0.04
Video has been playing uninterrupted for 30 seconds	1.04	2.21	0	0.00	-0.07
Seek within a video	1.38	6.68	0	0.01	-0.03
Toggle video caption on/off	0.00	0.11	0	0.00	0.00
Reach end of video	0.14	0.38	0	0.00	-0.03
Start a previous TYS or CYU assessment	0.01	0.13	0	0.00	0.01
Answer a previous item	1.46	1.86	1	0.05	-0.21
Go back to a previous item	0.06	0.36	0	0.01	0.07
Finish an assessment	0.93	0.97	1	0.02	0.11
Review a correctly answered item	0.01	0.17	0	0.00	0.02
Review an incorrectly answered item	1.10	1.94	0	0.29	0.08
Review solution video for a specific item	0.00	0.05	0	0.00	0.00
Review topic video for a specific quiz	0.00	0.02	0	0.00	0.00
Leave the assessment environment	0.67	0.88	1	0.01	-0.05
Navigate to the discussion board	0.81	0.88	1	0.14	-0.13
Load more entries on the discussion board	0.00	0.05	0	0.00	0.00
Make a post on the discussion board	0.00	0.01	0	0.00	0.00
Search on the discussion board	0.00	0.01	0	0.00	-0.01
Watch bio video for a tutor	0.00	0.08	0	0.00	-0.01
Karma awarded	0.00	0.01	0	0.00	0.00
Quiz Difficulty				0.22	-0.27
Student Ability				0.22	0.22

completed with the mirt R package [14]. For initial student ability estimates (where students do not have prior quiz data), student scores from the prior-year Florida State Assessment (FSA) standardized test were used. To represent general student ability, we calculated the average of student ability across all 10 sections. Given the way the item parameters were scaled (see [65]), the average student ability across the sections can be thought of as an average location in the standard unit normal distribution across sections. Both domain specific ability and general student ability were updated after the completion of each quiz. The average student ability using these estimates was 0.07 (IQR = 1.19). To represent quiz difficulty, we averaged the established difficulty scores (β) of the three quiz items. The average difficulty score was -0.07 (IQR = 0.91). The distribution of these estimates can be found in Figure 3

2.4 Dataset

We considered assessments taken over the 2018-2019 school year by 48,181 students in the state of Florida studying Algebra 1 and ranging from grades 6 to 12. We define a quiz as a unique set of individual items where the order of the items does not matter. In total, our data set includes 724,910 completed attempts of 1,065 unique quizzes. On average, each student attempted 15.05 quizzes (SD = 24.28) over the school year and the average quiz score was 1.59 (SD = 1.13) out of a possible score of 3.

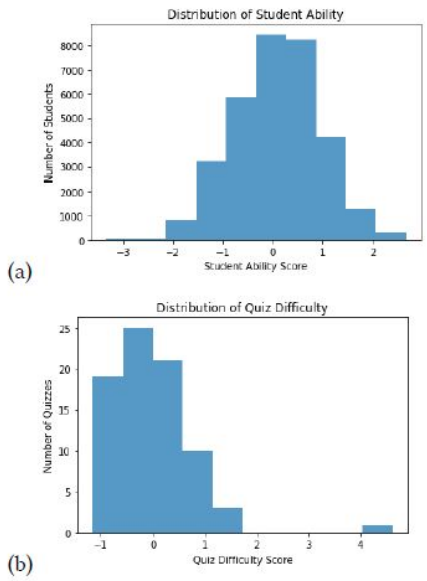

Figure 3: Distributions for (a) end-of-course estimated student ability (N = 32,685 students) and (b) quiz difficulty (N = 79 unique quizzes).

Since students can review the correct answers and retake each quiz for a better score, we focus for the rest of the paper on the *first attempt* of each student on a particular quiz. Of the total quiz attempts, 416,239 (57%) were the first attempt. Finally, we only considered quizzes that were associated with complete IRT data (discussed above), which produced a final data set of 210,190 quizzes. In the analyzed data set, each student had an average of 6.42 first attempts (SD = 8.79) with a first attempt score of 1.15/3 (SD = 1.07).

2.5 Machine Learning Procedures

In addition to the data cleaning noted above, we also removed any instances where no actions preceded the start of a quiz. This could occur, for instance, if a student remains logged into the platform overnight. As a result of this cleaning process, our final dataset consisted of 210,020 quiz attempts.

For the prediction task, we chose to use regression rather than classification since the data are not nominal categories. We trained a Random Forest regression model to predict a student's quiz score (ranging from 0 to 3). We trained these models using scikit-learn [49] and 5-fold student-level nested cross-validation; for each fold, the instances for each student were included in *either* the training or testing set. This practice reduces overfitting and promotes generalizability to new students. Additionally, we tuned our hyperparameters using a grid search; within each of the five iterations, the training set was split into three folds for validation. For each of these inner folds, a model was fit and scored using every combination of hyperparameters and tested on the held-out fold. The scores for each parameter combination across each validation fold were averaged, and the hyperparameters that resulted in the lowest mean squared error were preserved. A model was then fit on the full training set using these best parameters, and predictions were made on the test fold. These predictions were then pooled over the 5 test folds before accuracy metrics were computed. We considered 100, 300, and 500 estimators and a maximum tree depth of 6, 8, 10, 15, or none.

3 RESULTS

We evaluated the performance of each model using Pearson correlations (r) between actual and predicted quiz scores pooled over the five folds. The results are summarized in Table 2

3.1 Baseline Models

We compared our model performance against two random baselines. We first shuffled the true quiz scores (N = 210,020) across the entire dataset, which breaks the dependencies with the original features. We then trained a model using this shuffled data, which yielded a Pearson r = 0.00. Additionally, we shuffled the true quiz scores within students, removing any students with fewer than 5 quiz attempts (N = 171,105) as any fewer would not yield a meaningful baseline. A model trained on this shuffled data yielded a Pearson r = 0.24, which suggests it is learning an approximation of student ability as some students have a higher distribution of scores. Finally, we generated predicted scores using the 2PL framework, which considers estimated student ability, item difficulty, and item discrimination. These predicted scores yielded a correlation of 0.16 with the observed quiz scores.

3.2 Activity Feature Models

Our first question was whether we could prospectively predict quiz scores using only activity features. We initially trained a predictive model using the count of 22 activity features (N = 210,020, described in Table 1) in the 3 minutes leading up to the start of a quiz. This model achieved a Pearson r = 0.42. Next, we examined the influence of previous quiz activity on the current quiz (RQ2). This question is particularly interesting because some of the most frequent actions leading up to the quiz are related to taking another assessment. Since taking assessments is a form of retrieval practice (the behavior we want to promote), we anticipated that previous assessment activity in the same session would be associated with improved performance on the current quiz. To test this question, we first trained a model using the 13 activity features in Table 1 that were unrelated to taking assessments, including watching videos, interacting with the discussion wall, and general Algebra Nation onboarding activities. This model achieved an r of 0.33. Next, a model trained on the 9 features in Table 1 related to prior quiz activity (N = 210,020) yielded a Pearson r of 0.41. These results highlight the importance of previous retrieval practice on current assessment performance. It is important to emphasize that our data set only uses a student's first attempt at a quiz. Therefore, these results are not from students retaking a quiz for a higher grade; that is, activity from an entirely different quiz is a powerful predictor of future quiz performance.

3.3 IRT Models

The previous models showed some success in predicting student quiz scores only using simple activity features. These features are highly variable and change between student sessions on the Algebra Nation platform. This leads to the question, how much of a student's performance is consistent between learning sessions (RQ3)? We used the estimates of IRT features (quiz difficulty and student ability, N = 210,020) from the 2PL framework in our Random Forest model, which achieved a Pearson r of 0.34. This model differs from the baseline which predicts scores directly from the 2PL framework because it takes into account data from other students in the training set, rather than just the current estimates of student ability and quiz difficulty.

3.4 Combined Model

Finally, we investigated whether combining information from these different sources could improve our predictive performance. Accordingly, we trained a predictive model using the 22 activity features as well as quiz difficulty and student ability (24 features in all). We found that the combined feature model achieved the highest accuracy with a Pearson r = 0.53 (N = 210,020). This approach shows the promise of using both session-specific information as well as information about overall student ability and quiz difficulty, which do not depend on the particular learning session. The distribution of predicted and actual scores is in Figure 4

3.5 Different Session Lengths

We additionally considered how our predictive models are influenced by different session lengths (RQ4). Specifically, our question was whether we would be able to predict student quiz performance

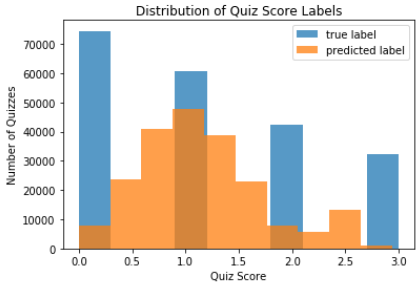

Figure 4: Distribution of true and predicted labels from the combined model (N=210,020)

Table 2: Performance comparison between the 3-minute models and random baseline (N = 210,020 except for the second baseline, N = 171,105).

Features	Pearson *r*
Baseline (shuffle across students)	0.00
Baseline (shuffle within students)	0.24
Baseline (IRT-generated scores)	0.16
All Activity Features	0.42
Non-quiz Activities	0.33
Only Prior Quiz Activities	0.41
IRT Features	0.34
Combined Features	0.53

using a longer session window. To test this question, we constructed different data sets that counted activity features in 5, 7, and 9-minute windows preceding the start of a quiz. As discussed above, we removed outliers by clipping each action to a 20-count maximum per one-minute interval. This allows for a higher maximum activity count for longer time windows.

The data sets constructed from longer session windows contained fewer instances than the original 3-minute data set because there are fewer sessions that contain activity over the longer length of time. Since there were different numbers of instances in each data set, we sampled each data set to contain the same number of instances across window sizes (N = 188,601). We trained the combined feature model on the various session windows and found no major difference as we increase the size of the session window; longer windows saw slightly lower scores, with the 9-minute window scoring 0.01 worse than the original model.

3.6 Feature Importance

We investigated the relative importance of individual features from the combined 3-minute window model (RQ5). For the Random Forest model, importance values are normalized to sum to 1, where higher values indicate more important features. This model captures nonlinearity and interactivity among features and is more difficult to interpret. Therefore for interpretability, we also used a cross-validated linear regression model predicting quiz scores using the 24 features in the combined model, which yielded a correlation of 0.36. We report standardized coefficients, where larger absolute values indicate a larger effect.

The two models generally aligned in terms of the most important features. The most important feature of the Random Forest model was reviewing an incorrect item on a previous assessment, with an importance of 0.29. In the linear regression model, this feature had a coefficient of 0.08, which means that reviewing incorrect items on previous quizzes was associated with higher scores on the current quiz. This is in line with previous research which shows retrieval practice promotes knowledge acquisition which can be transferred to new contexts [54]. Quiz difficulty and student ability were the next most important features in the Random Forest model, with an importance of 0.22. In the linear regression, quiz difficulty had a coefficient of -0.27, which is in line with the idea that more difficult quizzes are associated with lower scores (especially since we are considering first attempts here). Student ability had a coefficient of 0.22, which shows that higher performance on other quizzes in the platform is associated with higher first-attempt scores on subsequent quizzes; neither of these patterns are surprising. Loading the discussion board had an importance of 0.14 in the Random Forest model and a standardized linear coefficient of -0.13, which indicates that more visits to the discussion board is associated with lower quiz scores. This may indicate that students are visiting the discussion board simply to look up answers rather than thinking critically about solving the problem. Finally, answering a previous assessment item had an importance of 0.05 in the Random Forest model and a coefficient of -0.21, which indicates an interesting trend that answering more items on previous quizzes is associated with lower scores on the current quiz. The remaining features had significantly smaller importance in both models.

4 DISCUSSION

As an increased amount of learning is happening online, it is important that we promote effective learning strategies within these environments since students are poor at self-regulating their learning when left to their own devices [18]. Accordingly, our long-term goal is to nudge students towards retrieval practice (taking quizzes) when such a strategy would be beneficial in terms of learning, motivation, and engagement. As a step in this direction, we focused on developing a model that prospectively predicts student performance on short, formative quizzes in Algebra Nation. In the remainder of this section, we discuss our main findings, applications of our results, limitations, and future work.

4.1 Main Findings

Our overall finding is that our models were quite accurate (student-level cross-validated correlations of 0.53 or 28% of the variance) in prospectively modeling performance on an upcoming assessment using interaction patterns and past performance alone. Turning to our research questions, our first question addressed was how the immediate learning session can predict a student's performance on a quiz. We found that models that simply summed counts of generic activity features leading up to a quiz achieved a correlation of 0.42 in predicting quiz performance, easily outperforming two chance baselines. We then investigated the influence of prior quiz activity on the current quiz performance. We anticipated that participating

in more active learning strategies (e.g., retrieval practice) compared to passive learning strategies (watching videos) would be associated with better performance on assessments. Models trained on non-quiz features and only quiz features achieved correlations of 0.33 and 0.41, respectively, confirming our hypothesis.

Next, we considered how IRT features, proxies of student ability and quiz difficulty, can predict quiz performance. A model trained on these features achieved a correlation of 0.34, which is impressive since it does not take into account any activity that is unique to the particular learning session. By combining these features with the activity features, our model achieved a correlation of 0.53, which is the best of any of the conditions we tested.

We then asked whether considering longer session windows would improve our predictive performance since they are using more information about student learning activity. In almost all cases, we found that increasing the session window had no notable effect on model performance.

Finally, we investigated the relative importance of the features in our best-performing model, which combined all features over a 3-minute context window. Unsurprisingly, we found that the IRT features of quiz difficulty and student ability were negatively and positively associated with quiz score, respectively. An important new finding was that reviewing incorrect items was an important positive predictor of quiz score. Additionally, loading the discussion board was a negative predictor of quiz score. Since reviewing mistakes is a more active learning strategy than reading a discussion board (potentially to look up answers), these results align with previous work in self-regulated learning. The one important feature that is perhaps unintuitive is that answering items on a previous quiz was negatively associated with quiz score. This may be a result of cognitive fatigue; since we are considering only a 3-minute window before the start of a quiz, answering more items in the preceding window may deplete cognitive resources. But this speculative finding warrants more follow-up research.

4.2 Applications

The key application of this work is to integrate our model into the Algebra Nation platform so that it may be used as part of a larger intervention to promote retrieval practice as a study strategy. Such a model would accumulate data over overlapping activity windows and provide a real-time measure of prospective quiz performance. When the predicted score is sufficiently high, the Algebra Nation platform could suggest that students practice their skills by attempting a short quiz. Not only will this promote the use of retrieval practice as an effective learning strategy, but students will be able to validate their mastery of the course content and progress through the course material more efficiently.

We must be careful in how these recommendations are presented to students given that the model is not perfect and is not likely to ever be. For instance, if the system suggests that a student is ready to attempt a quiz and they do not perform well, this can have negative effects on their future motivation and trust in the intervention system. Thus, interventions should be designed to 'fail-soft' – there should be no negative impact on the student if the intervention is incorrectly delivered. For example, quiz recommendations can be based on the model's confidence and students should have an opportunity to opt-out of a quiz if they do not feel they are ready to attempt it. There is also the potential for using reinforcement learning techniques to learn an optimal policy for when to deliver a quiz. Finally, predicted quiz performance should not be used to dissuade a student from attempting a quiz; not only can retrieval practice be beneficial regardless of a student's score, but withholding quizzes can lead to a lack of feedback and impair student motivation.

Researchers considering applications of this work should interpret our findings in the context of the content, internal structure, and criterion-related validity evidence underlying the quiz item data. While continued development and validity studies are always needed, the assessment items in this work were supported by three core areas of validity evidence as defined by the *Standards for Educational and Psychological Testing* [1]. It is uncertain if the studied method for predicting quiz performance would uphold if used with measurement data lacking such validity evidence.

4.3 Limitations and Future Work

Like all research, ours has limitations. One such limitation is that we considered windows of activity immediately leading up to the start of a quiz. As such, we do not have information on how the model would perform without this orienting point. Future work should consider predicting student performance from the beginning of a learning session.

Our models were also relatively simple, consisting of only action counts and using Random Forest regressors. This was done as a first step towards this problem, but future work should examine whether performance can be improved by incorporating sequences of actions, latent action features, and deep learning methods. Similarly, though a strength of this work was the use of content-free features, this also presents a limitation. Such a feature set operates at a higher level of abstraction, which may aid generalizability at the cost of accuracy. In future work, we will examine the tradeoff between generalizability and accuracy by contrasting the content-free features used here with content-specific features such as identifiers for videos and quizzes.

In this work, we examined within one online learning platform, specifically aimed at mathematics. It is unclear at this time whether this work will generalize to additional domains, such as language learning or history. In future work, we hope to examine how this approach can be used in other domains and perhaps even other platforms (such as in [11]).

5 CONCLUDING REMARKS

In the classroom, teachers can support student learning through balancing passive (listening to a lecture or video) and active (working in groups, retrieval practice) learning strategies. In an online learning environment, students receive little to no guidance or feedback on how they should be spending their time. Timely interventions have the potential to help students navigate online learning environments more efficiently by encouraging use of effective learning strategies. In this work, we investigated whether we could predict student performance with sufficient accuracy to support such an approach. We did this by developing a predictive model of student performance on short, formative assessments in an online mathematics learning platform. We found that this model was most

successful when using both student activity in the immediate context as well as information on overall item difficulty and previous performance of the student on other algebra tests and quizzes. The models' overall performance was moderately high in that it explained 28% of the variance in subsequent quiz performance. Future work can investigate how suggestions based on these predictions can impact student behavior and learning outcomes.

ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305C160004 and Intel Research. The opinions expressed are those of the authors and do not represent views of the funding agencies.

REFERENCES

[1] (AERA), A.E.R.A. *et al.* 2014. *Standards for educational and psychological testing*.

[2] Adamopoulos, P. 2013. What makes a great mooc? an interdisciplinary analysis of student retention in online courses. *Thirty Fourth International Conference on Information Systems*. 2013, (2013), 1–21. DOI:https://doi.org/10.1145/1164394.1164397.

[3] Adesope, O.O. *et al.* 2017. Rethinking the use of tests: a meta-analysis of practice testing. *Review of Educational Research*. 87, 3 (2017), 659–701. DOI:https://doi.org/10.3102/0034654316689306.

[4] Agudo-Peregrina, Á.F. *et al.* 2014. Can we predict success from log data in vles? classification of interactions for learning analytics and their relation with performance in vle-supported f2f and online learning. *Computers in Human Behavior*. 31, (Feb. 2014), 542–550. DOI:https://doi.org/10.1016/j.chb.2013.05.031.

[5] Akyol, Z. and Garrison, D.R. 2011. Assessing metacognition in an online community of inquiry. *Internet and Higher Education*. 14, 3 (2011), 183–190. DOI:https://doi.org/10.1016/j.iheduc.2011.01.005.

[6] Ashenafi, M.M. *et al.* 2015. Predicting students' final exam scores from their course activities. *2015 IEEE Frontiers in Education Conference (FIE)* (Oct. 2015), 1–9. DOI:https://doi.org/10.1109/FIE.2015.7344081.

[7] Azevedo, R. *et al.* 2009. MetaTutor: a metacognitive tool for enhancing self-regulated learning. *AAAI Fall Symposium - Technical Report*. FS-09-02, (2009), 14–19.

[8] Azevedo, R. *et al.* 2010. Self-regulated learning with metatutor: advancing the science of learning with metacognitive tools bt - new science of learning: cognition, computers and collaboration in education. M.S. Khine and I.M. Saleh, eds. Springer New York. 225–247.

[9] Baker, R.S.J. d. *et al.* 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing bt - intelligent tutoring systems. (Berlin, Heidelberg, 2008), 406–415.

[10] Birnbaum, A. 1968. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*. F.M. Lord and M.R. Novick, eds. Addison-Wesley. 367–472.

[11] Boyer, S. and Veeramachaneni, K. 2015. Transfer learning for predictive models in massive open online courses bt - artificial intelligence in education. (Cham, 2015), 54–63.

[12] Brinton, C.G. and Chiang, M. 2015. MOOC performance prediction via clickstream data and social learning networks. *Proceedings - IEEE INFOCOM*. 26, (2015), 2299–2307. DOI:https://doi.org/10.1109/INFOCOM.2015.7218617.

[13] Brown, A. and Croudace, T.J. 2014. Scoring and estimating score precision using irt. *Handbook of item response theory modeling: Applications to typical performance assessment*. S.P. Reise and D.A. Revicki, eds. Routledge/Taylor & Francis Group. 307–333.

[14] Chalmers, R.P. 2012. Mirt: a multidimensional item response theory package for the r environment. *Journal of Statistical Software*. 48, 6 (2012). DOI:https://doi.org/10.18637/jss.v048.i06.

[15] Chaturvedi, R. and Ezeife, C.I. 2017. Predicting student performance in an its using task-driven features. *2017 IEEE International Conference on Computer and Information Technology (CIT)* (Aug. 2017), 168–175. DOI:https://doi.org/10.1109/CIT.2017.34.

[16] Chen, C.-M. *et al.* 2005. Personalized e-learning system using item response theory. *Computers & Education*. 44, 3 (Apr. 2005), 237–255. DOI:https://doi.org/10.1016/j.compedu.2004.01.006.

[17] Chi, M.T.H. and Wylie, R. 2014. The icap framework: linking cognitive engagement to active learning outcomes. *Educational Psychologist*. 49, 4 (Oct. 2014), 219–243. DOI:https://doi.org/10.1080/00461520.2014.965823.

[18] Coffrin, C. *et al.* 2014. Visualizing patterns of student engagement and performance in moocs. *Proceedins of the Fourth International Conference on Learning Analytics And Knowledge - LAK '14* (New York, New York, USA, 2014), 83–92. DOI:https://doi.org/10.1145/2567574.2567586.

[19] Corbett, A.T. and Anderson, J.R. 1995. Knowledge-tracing: modeling the acquisition of procedural knowledge. *User Modeling and User Adopted Interaction*. 4, (1995), 253–278. DOI:https://doi.org/10.1007/BF01099821.

[20] Dalipi, F. *et al.* 2018. MOOC dropout prediction using machine learning techniques: review and research challenges. *2018 IEEE Global Engineering Education Conference (EDUCON)* (2018), 1007–1014. DOI:https://doi.org/10.1109/EDUCON.2018.8363340.

[21] Doan, T.-N. and Sahebi, S. 2019. Rank-based tensor factorization for student performance prediction. *The 12th International Conference on Educational Data Mining* (2019), 288–293.

[22] Dwork, C. *et al.* 2012. Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference* (2012), 214–226. DOI:https://doi.org/10.1145/2090236.2090255.

[23] Elbadrawy, A. *et al.* 2015. Collaborative multi-regression models for predicting students' performance in course activities. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15* (New York, New York, USA, 2015), 103–107. DOI:https://doi.org/10.1145/2723576.2723590.

[24] Elbadrawy, A. *et al.* 2016. Predicting student performance using personalized analytics. *Computer*. 49, 4 (2016), 61–69. DOI:https://doi.org/10.1109/MC.2016.119.

[25] Emerson, A. *et al.* 2019. Predicting early and often: predictive student modeling for block-based programming environments. *The 12th International Conference on Educational Data Mining* (2019), 39–48.

[26] Hutt, S. *et al.* 2019. Time to scale: generalizable affect detection for tens of thousands of students across an entire school year. *2019 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)* (2019). DOI:https://doi.org/10.1145/3290607.3300726.

[27] Jacobson, M.J. 2008. A design framework for educational hypermedia systems: theory, research, and learning emerging scientific conceptual perspectives. *Educational Technology Research and Development*. 56, 1 (2008), 5–28. DOI:https://doi.org/10.1007/s11423-007-9065-2.

[28] Jensen, E. *et al.* 2019. Generalizability of sensor-free affect detection models in a longitudinal dataset of tens of thousands of students. *The 12th International Conference on Educational Data Mining* (2019), 324–329.

[29] Johnson, B.C. and Kiviniemi, M.T. 2009. The effect of online chapter quizzes on exam performance in an undergraduate social psychology course. *Teaching of Psychology*. 36, 1 (2009), 33–37. DOI:https://doi.org/10.1080/00986280802528972.

[30] Karpicke, J.D. *et al.* 2009. Metacognitive strategies in student learning: do students practise retrieval when they study on their own? *Memory*. 17, 4 (Apr. 2009), 471–479. DOI:https://doi.org/10.1080/09658210802647009.

[31] Koedinger, K.R. *et al.* 2015. Learning is not a spectator sport: doing is better than watching for learning from a mooc. *L@S 2015 - 2nd ACM Conference on Learning at Scale*. (2015), 111–120. DOI:https://doi.org/10.1145/2724660.2724681.

[32] Mao, Y. *et al.* 2018. Deep learning vs. bayesian knowledge tracing: student models for interventions. *Journal of Educational Data Mining*. 10, 2 (2018), 28–54.

[33] Mao, Y. *et al.* 2019. One minute is enough: early prediction of student success and event-level difficulty during a novice programming task. *Proceedings of the International Conference on Educational Data Mining* (2019), 119–128.

[34] Maris, G. and Bechger, T. 2009. On interpreting the model parameters for the three parameter logistic model. *Measurement: Interdisciplinary Research & Perspective*. 7, 2 (May 2009), 75–88. DOI:https://doi.org/10.1080/15366360903070385.

[35] Meier, Y. *et al.* 2016. Predicting grades. *IEEE Transactions on Signal Processing*. 64, 4 (Feb. 2016), 959–972. DOI:https://doi.org/10.1109/TSP.2015.2496278.

[36] Milligan, C. *et al.* 2013. Patterns of engagement in massive open online courses. *Journal of Online Learning with Technology (Special Issue on MOOCs)–Under Review*. 9, 2 (2013), 149–159.

[37] Moos, D.C. 2014. Setting the stage for the metacognition during hypermedia learning: what motivation constructs matter? *Computers and Education*. 70, (2014), 128–137. DOI:https://doi.org/10.1016/j.compedu.2013.08.014.

[38] Moreira, B.F.T. *et al.* 2019. Retrieval practice in classroom settings: a review of applied research. *Frontiers in Education*. 4, February (2019). DOI:https://doi.org/10.3389/feduc.2019.00005.

[39] Moreno-Marcos, P.M. *et al.* 2019. Prediction in moocs: a review and future research directions. *IEEE Transactions on Learning Technologies*. 12, 3 (2019), 384–401. DOI:https://doi.org/10.1109/TLT.2018.2856808.

[40] Mulligan, N.W. and Picklesimer, M. 2016. Attention and the testing effect. *Journal of Experimental Psychology: Learning Memory and Cognition*. 42, 6 (2016), 938–950. DOI:https://doi.org/10.1037/xlm0000227.

[41] Niaki, S.A. *et al.* 2019. Investigating the usage patterns of algebra nation tutoring platform. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19* (New York, New York, USA, 2019), 481–490. DOI:https://doi.org/10.1145/3303772.3303788.

[42] Palmatier, R.A. and Bennett, J.M. 1974. Notetaking habits of college students. *Journal of Reading*. 18, 3 (Mar. 1974), 215–218.

[43] Pardos, Z.A. *et al.* 2013. Adapting bayesian knowledge tracing to a massive open online course in edx. *Proceedings of the 6th International Conference on Educational Data Mining, EDM 2013*. (2013).

[44] Pardos, Z.A. *et al.* 2012. The sum is greater than the parts: ensembling models of student knowledge in educational software. *ACM SIGKDD Explorations Newsletter*. 13, 2 (May 2012), 37. DOI:https://doi.org/10.1145/2207243.2207249.

[45] Pardos, Z.A. *et al.* 2010. Using fine-grained skill models to fit student performance with bayesian networks. CRC Press.

[46] Pardos, Z.A. and Heffernan, N.T. 2011. KT-idem: introducing item difficulty to the knowledge tracing model. *International Conference on User Modeling, Adaptation, and Personalization* (Berlin, Heidelberg, 2011), 243–254.

[47] Partchev, I. 2009. 3PL: a useful model with a mild estimation problem. *Measurement: Interdisciplinary Research & Perspective*. 7, 2 (May 2009), 94–96. DOI:https://doi.org/10.1080/15366360903117046.

[48] Pavlik, P.I. *et al.* 2009. Performance factors analysis –a new alternative to knowledge tracing. *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling* (NLD, 2009), 531–538.

[49] Pedregosa, F. *et al.* 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*. 12, (2011), 2825–2830. DOI:https://doi.org/10.1007/s13398-014-0173-7.2.

[50] Pellas, N. 2014. The influence of computer self-efficacy, metacognitive self-regulation and self-esteem on student engagement in online learning programs: evidence from the virtual world of second life. *Computers in Human Behavior*. 35, (2014), 157–170. DOI:https://doi.org/10.1016/j.chb.2014.02.048.

[51] Price, M.J. *et al.* 2018. The role of negative emotions and emotion regulation on self-regulated learning with metatutor. *Proceedings of the International Conference on Intelligent Tutoring Systems* (Cham, 2018), 170–179.

[52] Rajendran, R. and Biswas, G. 2016. Modeling learners' metacognitive skills in open ended learning environments. *ICCE 2016 - 24th International Conference on Computers in Education: Think Global Act Local - Workshop Proceedings*. (2016), 407–412.

[53] Ramos, C. and Yudko, E. 2008. "Hits" (not "discussion posts") predict student success in online courses: a double cross-validation study. *Computers and Education*. 50, 4 (2008), 1174–1182. DOI:https://doi.org/10.1016/j.compedu.2006.11.003.

[54] Roediger, H.L. and Butler, A.C. 2011. The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*. 15, 1 (2011), 20–27. DOI:https://doi.org/10.1016/j.tics.2010.09.003.

[55] Roediger, H.L. and Karpicke, J.D. 2006. Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science*. 17, 3 (2006), 249–255. DOI:https://doi.org/10.1111/j.1467-9280.2006.01693.x.

[56] Sahebi, S. and Brusilovsky, P. 2018. Student performance prediction by discovering inter-activity relations. *Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018* (2018).

[57] Sinha, T. and Cassell, J. 2015. Connecting the dots: predicting student grade sequences from bursty mooc interactions over time. *Proceedings of the Second (2015) ACM Conference on Learning @ Scale* (New York, NY, USA, 2015), 249–252.

[58] Szpunar, K.K. *et al.* 2013. Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*. 110, 16 (Apr. 2013), 6313–6317. DOI:https://doi.org/10.1073/pnas.1221764110.

[59] Szpunar, K.K. *et al.* 2008. Testing during study insulates against the buildup of proactive interference. *Journal of experimental psychology. Learning, memory, and cognition*. 34, 6 (Nov. 2008), 1392–9. DOI:https://doi.org/10.1037/a0013082.

[60] Thaler, R.H. and Sunstein, C.R. 2008. *Nudge: improving decisions about health, wealth and happiness*. Yale University Press.

[61] Wainer, H. and Mislevy, R.J. 1990. Item response theory, item calibration, and proficiency estimation. *Computer adaptive testing: A primer*. H. Wainer, ed. Lawrence Erlbaum. 65–102.

[62] Weinstein, Y. *et al.* 2011. Testing protects against proactive interference in face-name learning. *Psychonomic bulletin & review*. 18, 3 (Jun. 2011), 518–23. DOI:https://doi.org/10.3758/s13423-011-0085-x.

[63] Winne, P.H. and Hadwin, A.F. 1998. Studying as self-regulated learning. *Metacognition in educational theory and practice*. D.J. Hacker *et al.*, eds. Erlbaum. 277–304.

[64] Wissman, K.T. *et al.* 2011. The interim test effect: testing prior material can facilitate the learning of new material. *Psychonomic bulletin & review*. 18, 6 (Dec. 2011), 1140–7. DOI:https://doi.org/10.3758/s13423-011-0140-7.

[65] Xue, K. *et al.* 2020. Semi-supervised learning method for adjusting biased item difficulty estimates caused by nonignorable missingness under 2pl-irt model. *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020* (2020), 715–719.

[66] Yudelson, M. V *et al.* 2013. Individualized bayesian knowledge tracing models bt - artificial intelligence in education. (Berlin, Heidelberg, 2013), 171–180.

[67] Zhang, Y. *et al.* 2020. The relationship between confusion and metacognitive strategies in betty's brain. *The 10th International Learning Analytics and Knowledge* (Frankfurt, Germany, 2020).

[68] Zou, J. and Schiebinger, L. 2018. AI can be sexist and racist — it's time to make it fair. *Nature*. (2018). DOI:https://doi.org/10.1038/d41586-018-05707-8.