Adaptation of the Multi-Concept Multivariate Elo Rating System to Medical Students' Training Data

Erva Nihan Kandemir, Jill-Jênn Vie, Adam Hegel Sanchez Ayte, Olivier Palombi, Franck Ramus

Adaptation of the Multi-Concept Multivariate Elo Rating System to Medical Students' Training Data

Erva Nihan Kandemir
erva.nihan.kandemir@ens.psl.eu
École Normale Supérieure, PSL University, CNRS
Paris, 75005, France

Jill-Jênn Vie
vie@jill-jenn.net
Soda team, Inria Saclay
France

Adam Sanchez-Ayte
adam.sanchez@uness.fr
Université Numerique en Santé et Sport (UNESS)
France

Olivier Palombi
olivier.palombi@univ-grenoble-alpes.fr
Université Grenoble Alpes, Grenoble INP, CNRS, Inria, LIG
Grenoble, 38000, France

Franck Ramus
franck.ramus@ens.psl.eu
École Normale Supérieure, PSL University, EHESS, CNRS
Paris, 75005, France

ABSTRACT

Accurate estimation of question difficulty and prediction of student performance play key roles in optimizing educational instruction and enhancing learning outcomes within digital learning platforms. The Elo rating system is widely recognized for its proficiency in predicting student performance by estimating both question difficulty and student ability while providing computational efficiency and real-time adaptivity. This paper presents an adaptation of a multi-concept variant of the Elo rating system to the data collected by a medical training platform—a platform characterized by a vast knowledge corpus, substantial inter-concept overlap, a huge question bank with significant sparsity in user-question interactions, and a highly diverse user population, presenting unique challenges. Our study is driven by two primary objectives: firstly, to comprehensively evaluate the Elo rating system's capabilities on this real-life data, and secondly, to tackle the issue of imprecise early-stage estimations when implementing the Elo rating system for online assessments. Our findings suggest that the Elo rating system exhibits comparable accuracy to the well-established logistic regression model in predicting final exam outcomes for users within our digital platform. Furthermore, results underscore that initializing Elo rating estimates with historical data remarkably reduces errors and enhances prediction accuracy, especially during the initial phases of student interactions.

CCS CONCEPTS

• **Human-centered computing** → **User models**; • **Applied computing** → **E-learning**.

KEYWORDS

knowledge tracing, Elo-based learning model, logistic regression

1 INTRODUCTION

Over the last decade, as a result of the increasing use of educational technology involving significant amounts of data and learning analytics, personalized learning has gained increasing popularity. This has led a number of research groups to study the adaptation of personalized learning into educational technologies, leveraging insights from learning analytics [5, 19]. Adaptive Learning Systems (ALSs) [18] achieve personalized learning experiences by using users' prior interactions and by adjusting the learning content to match individual preferences and requirements. Research consistently highlights the effectiveness of ALSs when compared to non-adaptive systems, resulting in improved learning outcomes [20, 21, 31, 33] and a positive impact on student motivation [3], engagement [22], and comprehension [4]. Today, prominent ALS platforms like Duolingo, and ALEKS deliver high-quality learning materials and personalized instruction to millions of users worldwide.

With all the benefits listed above, this adaptive method in online education requires effectively monitoring users' learning paths, a procedure called knowledge tracing. This knowledge-tracing process involves creating learner models based on student performance and interactions with the system to represent their ability levels and the difficulty of educational materials.

1.1 Background

Various models have been designed to monitor and predict students' evolving knowledge levels over time [8]. These models can be broadly classified into four categories: Markov process models [12, 16], logistic models [11, 20, 28, 32], deep knowledge tracing models [10, 15, 27, 29, 30], and rating systems [3]. The first three classes of these models are well-established and already extensively documented in existing literature [1]. Although they exhibit strong prediction capabilities when evaluating student performance, they are not without limitations, particularly when deploying them in online educational environments, where they often demand intricate parameter estimation and calibration procedures, typically relying on large datasets. This complexity can impede the development of adaptive systems, making them more challenging, time-consuming, and resource-intensive to create.

An intriguing alternative is the utilization of rating systems, offering a computationally more economical approach. Rating Systems, particularly the Elo Rating system [14], have been widely applied in educational technologies [7, 17, 23, 26]. Originally created for ranking chess players, the Elo rating system has been repurposed

for educational settings, treating users and learning materials as opponents. In the educational context, it predicts ratings for users and questions, serving as an assessment of user ability and question difficulty.

In this framework, each user u is associated with a global ability parameter denoted as θ_u. Similarly, for each question i, there exists a question parameter θ_i reflecting the difficulty level of that question. The probability of a user u correctly attempting a multiple-choice question i, denoted as $\Pr(a_{ui} = 1|\theta_u, \theta_i)$, can be expressed as a logistic function of the difference between the user and question parameters:

$$\Pr(a_{ui} = 1|\theta_u, \theta_i) = \sigma(\theta_u - \theta_i) = \frac{1}{1 + e^{-(\theta_u - \theta_i)}}$$

After a user u attempts question i, both the user's ability and the question's difficulty undergo updates that are proportional to the difference between the estimated probability and the actual outcome. These updates are defined by the following formulas for question difficulty (θ_i) and for user ability (θ_u):

$$\begin{aligned} \theta_i &:= \theta_i + K(\Pr(a_{ui} = 1|\theta_u, \theta_i) - a_{ui}) \\ \theta_u &:= \theta_u + K(a_{ui} - \Pr(a_{ui} = 1|\theta_u, \theta_i)) \end{aligned} \tag{1}$$

Here, a_{ui} represents the actual outcome of the attempt of the user u on question i, and K is a constant value that determines the degree of update sensitivity based on the user's most recent attempt. The choice of the constant K in the update rule plays a pivotal role in shaping estimation dynamics. If K is set too low, the estimation process progresses too slowly, leading to prolonged uncertainty in skill assessment and potential failure to reach correct values. Conversely, if K is set too high, the system is unstable, heavily influenced by recent attempts, and thus provides erratic evaluations.

In light of this formulation, the classical Elo rating system in education reveals its iterative nature, refining user and question parameters after each interaction.

While the Elo rating system does not provide statistically guaranteed estimations, in contrast to well-calibrated logistic models, numerous studies have explored the accuracy of the Elo rating system and compared its performance to state-of-the-art models. For instance, study [35] found that the IRT-Rasch model version, proportion correct, and the Elo rating system, increasingly correlate with the true difficulty parameter as sample size increases. Another study [24] compared Elo rating's question difficulty estimates with those obtained through the joint maximum likelihood method (JMLE) and observed nearly identical outcomes. Studies using simulated data [6, 25, 26] have also concluded that Elo rating systems perform similarly to the Rasch model, making them suitable for systems requiring real-time user knowledge adaptation without the need for extensive question pretesting on large sample sizes.

However, integrating the Elo rating algorithm into a real-time educational context presents challenges, especially in the initial stages when student abilities and question difficulties are unknown and are set to zero. Given the iterative nature of the model, these initial estimates are assumed to gradually correct themselves with each attempt. While starting the model from scratch is standard practice, to produce reliable estimates the system requires a substantial number of responses, typically at least 100 attempts [26]. Furthermore, uncertain initial estimates can exert a lasting influence on subsequent updates, potentially resulting in persistent inaccuracies. This issue is especially pronounced for questions and users with a limited number of attempts within the educational platform, as they may have fewer opportunities for correction.

1.2 Goals of the present study

In this study, we have two primary objectives: firstly, to assess whether the Elo rating system meets the requirements of a complex real-life scenario with specific challenges, and secondly, to address the issue of imprecise early-stage estimations in the online application of the Elo rating system. To mitigate the uncertainty associated with initial Elo rating estimations, we used data collected in the previous year.

In brief, our learning platform is open to all French medical students throughout their studies to provide training for their medical studies (about 8600 students per year over 10 years, with one important national exam at the end of the 6^{th} year). The challenges raised by this particular learning context are multiple:

- The knowledge corpus is huge, as it encompasses *all medical knowledge* taught in French universities.
- This corpus is structured into knowledge components that are themselves very large: 362 distinct topics or subcategories (themes/areas) of medical knowledge, spread over 31 medical specialties.
- The corpus of questions is also huge (~1,500,000 questions), such that a given student takes only a tiny fraction of available questions (usually drawn at random), almost never takes the same question twice, and such that a given question is only taken by a tiny fraction of students (outside exams). Thus, the matrix is extremely large and sparse.
- Use of the training platform is optional, with some students using it intensively on a daily basis, and others doing most of their training outside the platform.
- Students are based in 42 universities which cover the same curriculum from 1^{st} to 6^{th} year, but with different material and in a different order.

Despite these difficulties, the fact that all medical students from all universities can train and take exams on the same platform should make it possible and desirable to model their progress and use this modeling to provide them with an adaptive training program. Yet the first step is to show that a sufficiently reliable modeling of their progress is possible under the present conditions.

Thus, the main research question guiding this study is to examine the boundaries of the Elo rating system within a particularly challenging real-life scenario that encompasses various complexities. This evaluation involves a comparison of its accuracy against the widely accepted logistic regression model. Additionally, we seek to explore the potential advantages of initializing the Elo system with results obtained from logistic regression applied to data from preceding years.

2 METHOD

This study employed an observational research design, leveraging ecological data from the existing BNE (*Banque nationale d'entraînement*) digital learning platform.

2.1 BNE Platform

The BNE digital learning system serves as an online platform extensively utilized by over 8,800 medical students across all French universities annually. This platform is used by all medical faculties to administer exams. Exam questions are then added to the question bank, together with additional questions designed by professors for training purposes. This platform therefore holds a very large set of multiple-choice questions covering 31 medical specialties that are made available to students for training. For medical students, the platform is a valuable resource to prepare for the ECN (*Épreuves Classantes Nationales*) final exam. This final exam usually takes place in June of the sixth academic year and significantly influences the choice of students' medical specialization.

To enhance the pedagogical engagement of medical students on the platform, a notable feature allows learners to tailor their training experience. They can choose from various question types and medical specialties, enabling them to simulate and practice for a wide range of medical exams according to their preferences and needs.

2.2 BNE data set Overview

Within this section, we describe the BNE data set for the educational year 2020-2021, representing the most recent accessible data sourced from the BNE platform during our analysis. Additionally, we describe the usage patterns observed on the platform during this year, providing valuable insights into its structure and functionality.

From the BNE platform, direct access to the official ECN exam is unavailable. However, we do have access to a mock exam, typically conducted in mid-March (specifically on March 15th, 16th, and 17th, 2021, for the 2020-2021 academic year). This mock exam closely mimics the format of the actual ECN final exam. For the purpose of testing student prediction models on our complex data set, we focused our analysis on the educational year of 2020-2021, specifically targeting 6th-year users who participated in the mock final ECN exam. This selection aligns with the core objective of our study, which is to assess the models' performance through external validation using the mock final exam.

In the following sections, we will describe the data related to the six-month training period spanning from September 16, 2020 to March 14, 2021, leading up to the mock final exam, distinct from the data associated with the mock exam itself, on March 15th, 16th, and 17th, 2021.

2.2.1 Training Period data set. Table 1 offers a comprehensive overview of our training period data set's key characteristics, including the total number of *users* (8,616), *questions* (357,317), *medical specialties* (31), and *attempts* (26,772,424).

Additionally, the *specialty per question* variable indicates the average number of specialty tags associated with each question. Here, each of the 31 medical specialties serves as a distinct knowledge component (KC). These knowledge components are much larger than is usually defined in the literature. In this context, when a question i is tagged with a specific specialty s, the likelihood of correctly answering question i hinges upon the user's specialty-specific ability s. Conversely, we assess a user's ability in specialty s by evaluating their ability to successfully tackle questions tagged with the same specialty s. The table reveals that the average number of specialties associated with each question (1.58) is greater than 1, underscoring that certain questions require knowledge spanning multiple medical specialties for accurate responses. In this context, our data set can be characterized as a multi-knowledge component data set (or multi-specialty data set in the specific context of BNE). For a more in-depth exploration of this distribution, Figure 1 provides a detailed breakdown of the count of questions requiring varying numbers of specialties.

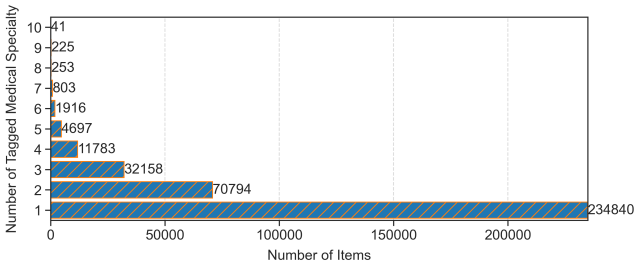

Figure 1: Number of questions that require knowledge on any given number of medical specialties.
The questions exhibit a spectrum of dependence on medical specialty knowledge for their solution. While a considerable portion of questions rely on ability in a single medical specialty, many questions require knowledge spanning multiple specialties.

The user-question sparsity in Table 1 indicates the proportion of missing values in the user-question interaction matrix. The table shows that our data set is substantial and exhibits significant sparsity (*Sparsity (user, question)* = 0.99). There are vastly more available questions than any user can take, and different users will take different questions (usually by random draw).

Lastly, the *Attempts per User* variable unveils the average number of attempts made by the same user on individual questions. This indicates whether students frequently revisit questions they have previously encountered. With a value of 1.05, it is evident that during the training period, students almost never re-attempt questions they have already attempted.

Figure 2 provides a detailed description of the platform's usage patterns across 31 available medical specialties during the 2020-2021 educational year. As mentioned earlier, the platform provides users with the option to create their own training sessions by selecting specific medical specialties and question types they want to study. This results in varying levels of popularity among the specialties. The figure reveals that while most students engage with questions from all specialties, there is considerable variability in both the quantity of available questions and the number of questions taken within each specialty.

Table 1: BNE data set Summary

Data	Period	Users	Questions	Specialties	Specialties per Question	Attempts	Sparsity (User, Question)	Attempts per User
Training Period data set	**16.09.2020–14.03.2021**	8,616	357,317	31	1.58	26,772,424	0.99	1.05
Mock Final Exam data set	**15–17.03.2021**	8,616	372	28	1.71	3,172,546	0.01	1

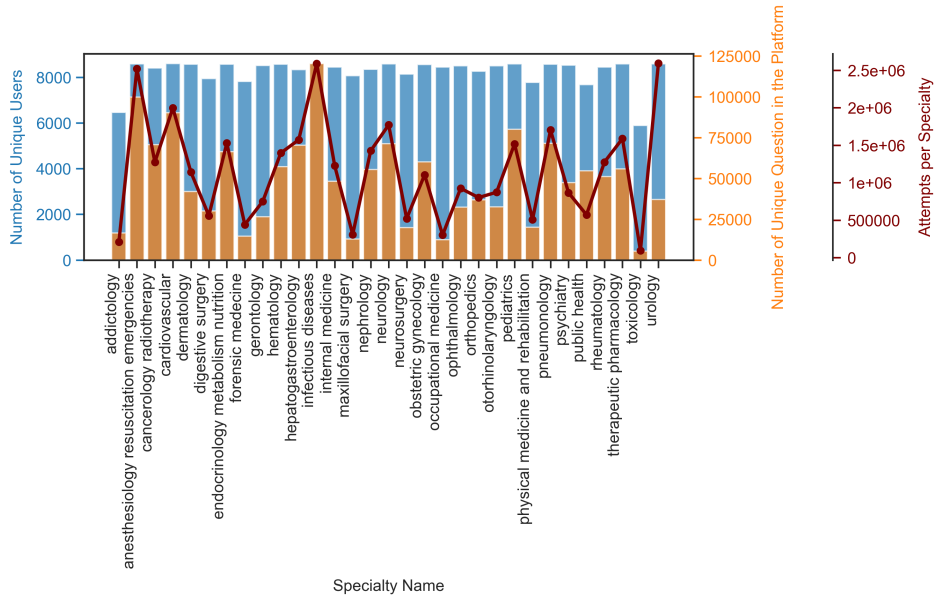

Figure 2: Overview of the use of the BNE Platform during the 2020-2021 educational year.
The blue bars represent the count of unique users per medical specialty in the data set. The orange bars represent the count of unique questions available in the platform for each specialty. In addition, the overlaid line plot illustrates 'Attempts per Specialty,' the total number of user attempts on questions within each specialty during the educational year 2020-2021.

2.2.2 Mock Final Exam data set. Table 1 also provides a comprehensive overview of the nature of the mock ECN final exam for the educational year 2020-2021, held on the 15th, 16th, and 17th of March 2021. The Mock final exam data set encompasses data from 8,616 users who took 372 questions spanning 28 distinct medical specialties (addictology, orthopedics, and toxicology were not included in the mock exam). Questions in this data set were associated with a mean number of 1.71 specialties, reflecting greater multidisciplinarity of questions than in the training period data set. The mock final exam data set recorded a total of 3,172,546 user interactions, highlighting a high level of user engagement, with a minimal sparsity value of 0.01, implying that almost all users attempted every question during the final exam. Additionally, the *Attempts per User* value of 1 indicates that users made only a single attempt at each question.

Here, it's worth noting that all the mock exam questions were entirely new, distinct from those encountered during the training period. Therefore, the data from the mock final exam does not provide a direct test of the knowledge of specific questions taken in the training period, but rather a test of students' ability to generalize what they have learned during courses and training to new questions in the same medical specialties, mirroring the format of the official ECN exam, which emphasizes generalization rather than memorization of specific knowledge.

2.3 Elo Rating: Model Extensions for Adaptation to the BNE data set

The standard iterative formulation of the Elo rating system, which computes user and question-related factors, has been previously

described in the related literature [6, 7, 17, 23–26]. To optimize its adaptability for educational contexts, the Elo rating system has undergone numerous extensions. In this section, we indicate the specific modifications we have applied to tailor the model to our BNE data set.

2.3.1 Incorporating the guessing behavior into the Elo rating system. In numerous studies employing the Elo rating as a prediction model for multiple-choice questions, researchers take into account the guessing rate when calculating the probability of correctness [24, 26].

In such instances, the probability of a user u attempting a multiple-choice question i with n_{opt} choices correctly, denoted as $\Pr(\text{correct} \mid \theta_u, \theta_i)$, can be expressed as follows:

$$\Pr(a_{ui} = 1 \mid \theta_u, \theta_i) = \Pr(\text{guessing} \mid n_{\text{opt}}) + \frac{1 - \Pr(\text{guessing} \mid n_{\text{opt}})}{1 + e^{-(\theta_u - \theta_i)}}$$

Within the BNE question pool, questions are divided into two main types: single- and multiple-answer questions. For single-selection questions (unique answer questions), $P(\text{guessing}|n_{\text{opt}})$ is straightforward, equating to $1/n_{\text{opt}}$. For multiple choice questions, $P(\text{guessing}|n_{\text{opt}})$ is the inverse of the sum of the possible ways to select any number of answers k from the available options:

$$\Pr(\text{guessing} \mid n_{\text{opt}}) = \frac{1}{\sum_{k=1}^{n_{\text{opt}}} \binom{n_{\text{opt}}}{k}}$$

2.3.2 Decreasing Uncertainty. The dynamics of the Elo rating system in educational settings are crucial for accurately assessing the skills and abilities of students and questions. The challenge lies in managing evolving uncertainties, which are inherently dynamic. When new students or questions are introduced to the platform, our information on their true abilities or difficulties is limited, meaning high uncertainty. Consequently, during this initial phase, it is essential for the model to make significant updates to its estimates. As more data accumulates, students engage in multiple attempts, and questions are extensively attempted by a set of students, and the model should naturally become more certain about its estimation of ability levels or difficulty levels. In such cases, the model should reduce the update parameter as confidence in the estimates grows.

In order to meet this challenge, recent applications of the Elo rating system in educational contexts [3] have replaced the fixed constant K in Equation 1 with a dynamic uncertainty function. This function, denoted as $U(n)$, is defined as:

$$U(n) = \frac{a}{1 + bn}$$

where a is the constant hyper-parameter determining the starting value; b is the constant hyper-parameter determining the slope of changes; n is the number of prior attempts of the user or question parameter.

The exact parameter values, as highlighted by [26], carry relatively less weight, as different choices for a and b tend to yield remarkably similar outcomes. In our case, we set $a = 1$ and $b = 0.5$ for both question and user attempts. These values were determined through an optimization process using grid search. However, it is important to mention that our model consistently delivered stable performance, and the precise choice of parameter values had only a negligible effect on the results.

Moreover, in keeping with the central aim of learner models, which is to effectively track shifts in user abilities, we have introduced a lower bound for the uncertainty function applied to user ability. By incorporating this lower bound of 0.03 into the user uncertainty function, we ensured that user ability updates persist even after a considerable number of attempts. With our current values of a and b, this lower bound applies after 65 attempts.

2.3.3 Multi-tag Knowledge Component Extension. As previously described, in our BNE data one question may be tagged by multiple specialties. To account for the ability of users in each of the tagged medical specialties separately we used the multi-concept extended version of the Elo rating introduced by [3]. The difference is that, instead of having only one global user ability parameter θ_u, we estimated user ability θ_{us} for each specialty s.It is important to note that, given the absence of information regarding the relative importance of tagged specialties for each question in the data, we adopted a straightforward approach as outlined in [3]. Specifically, we computed the mean ability λ_{ui} of student u on question i by averaging user u's abilities across all medical specialties $s_1, \ldots, s_\delta$ tagging question i, assigning equal weight to each specialty in this calculation.

$$\lambda_{ui} = \frac{1}{\delta} \sum_{k=1}^{\delta} \theta_{us_k}$$

Furthermore, not all specialties may have the same average difficulty level. In order to alleviate this, we define and estimate distinct parameters for question difficulty (d_i) and specialty difficulty (θ_s). We denote by μ_i the sum of question difficulty d_i and the average of difficulties of all skills $s_1, \ldots, s_\delta$ involved in question i:

$$\mu_i = d_i + \frac{1}{\delta} \sum_{k=1}^{\delta} \theta_{s_k}$$

Thus the probability of answering correctly becomes:

$$\Pr(a_{ui} = 1 \mid \lambda_{ui}, \mu_i) = p(\lambda_{ui}, \mu_i) \triangleq \sigma(\lambda_{ui} - \mu_i)$$

The update of the question difficulty d_i remains the same. However, now the update on the user skill parameters θ_{us} occurs on each tagged specialty separately, and the update for specialty difficulty θ_s follows a similar pattern as the updates for item difficulty:

$$\begin{aligned} d_i &:= d_i + U(n)\,(p(\lambda_{ui}, \mu_i) - a_{ui}) \\ \theta_{us} &:= \theta_{us} + U(n)\,(a_{ui} - p(\theta_{us}, d_i + \theta_s)) \\ \theta_s &:= \theta_s + U(n)\,(p(\theta_{us}, d_i + \theta_s) - a_{ui})\,. \end{aligned}$$

It's important to note that while updating θ_{us} and θ_s, the prediction formula operates at the specialty level for each tagged specialty, just like in [3], although the d_i update is based on question-level prediction.

By utilizing the Elo rating system along with the aforementioned extensions, it becomes possible to estimate three critical aspects: user ability in each specialty, questions' individual difficulty, and specialties' global difficulty.

2.4 Data Preparation Process

Before starting to train the Elo rating model and Logistic regression over the 2020-2021 data set, we performed a series of pre-processing steps on the combined data from the training period data set and

the mock ECN final exam data set. These steps were carried out in the following order:

- Removal of duplicated rows: 267 rows out of 29,900,533 were removed.
- Exclusion of questions without any tagged medical specialty: None removed (the data extraction process was already limited to questions with tagged specialties), but 30% lacked specialty tags initially.
- Exclusion of questions that are neither unique nor multiple-choice questions (open-answer questions): None removed.
- Binarization of question ratings (BNE has a more sophisticated rating scheme depending on the number of correct and incorrect answers ticked).
- Removal of users with fewer than 100 interactions: No questions or students were removed during this step. Since all students in the dataset had taken the ECN mock exam, they all had at least 100 attempts.
- Removal of questions with fewer than 100 interactions: 79.11% of the unique questions were removed.

As a result, our training period data set now consists of 22,294,780 attempts, made by 8,616 distinct users to 74,704 unique questions across 31 medical specialties. The mock ECN final exam data set includes 3,172,546 attempts. Within that data set, there are 372 unique questions taken by 8,616 users across 28 distinct medical specialties.

Figure 3 offers a visual depiction of the distribution of answers across students, questions, and specialties in both the training period and the mock final exam data sets after the pre-processing.

2.5 Information Encoding & Initialization of Elo Ratings via Logistic Regression Outputs

As previously mentioned, in the Elo rating system, initial estimates for both questions and users are typically set to 0, which can lead to high uncertainty. To address this, an alternative approach is to use the logistic regression outcomes of the previous year's data as informed initial values for initializing Elo ratings, rather than starting from scratch. With this approach, the Elo rating algorithm is anticipated to converge faster and more accurately, providing a "head start" that conserves computational resources and leads to more precise estimates.

To prepare the extensive BNE dataset for logistic regression modeling, we employed a one-hot vector encoding method inspired by [34]. This technique transformed each attempt in the original data set into a sparse vector containing all relevant information. In our adaptation of this approach, we aimed to closely align our logistic regression model with the principles of Elo rating estimations, while also incorporating all the aforementioned extensions we applied for the Elo rating system. To achieve this, we included the one-hot encoding of user-specialties interaction, question, and specialty for each attempt. With this approach, the logistic regression was able to estimate users' ability in each specialty, the difficulty of individual questions, and the overall difficulty of each specialty.

To implement the initialization approach, an essential step involves comparing the logistic regression model against the Elo rating system, utilizing data from the same year. This step aimed to ensure that, before employing the logistic regression model on the previous year's data and utilizing its outcomes for initialization purposes, the model aligned with the Elo framework, generating compatible and consistent results.

Subsequently, we applied the logistic regression model to the data from the 2019-2020 educational year, utilizing the outcome estimates as initial values for the Elo rating process applied to the 2020-2021 data. Our data set for the academic year 2019-2020 (spanning from September 15, 2019, to March 1, 2020) includes 400,774 distinct questions and 25,978 unique users. However, after filtering data to retain questions and users with enough attempts (cf. above), only 47,579 questions and 8,239 users were shared between 2019-2020 and 2020-2021.

As a result, we initialized the question difficulty and student ability values for the 2020-2021 academic year using the estimates obtained from the logistic regression model applied to the 2019-2020 data, whenever these values were available. In cases where values were not present in the 2019-2020 data set for a particular question or student, we initialized their 2020-2021 values to zero.

In order to allow the uncertainty function to be able to further update those initialized values, without entirely destabilizing the estimations, we set the initial number of previous attempts to 50, which seemed a reasonable compromise between the actual number of attempts (>100 which would make any update negligible) and 0 (which would underweight the previous history).

2.6 Performance Evaluation Metrics

In our performance evaluation, we examined the effectiveness of two variants of the Elo rating system on the training data set. One variant initialized all ability and difficulty values to 0, while the other initialized values based on logistic regression from the previous year. We also compared these Elo variants with logistic regression on the entire training data set. We used several key statistics, including Area Under the Curve (AUC), Root Mean Squared Error (RMSE), and Accuracy (ACC), to assess the prediction performance of these models first on the training period and second on the mock exam.

First, to comprehensively evaluate the prediction capabilities of the Elo rating system throughout its iterations, mirroring its real-world use within the platform, and understand the impact of initializing estimates via logistic regression, we compared these three models during the training period. We calculated the AUC, ACC, and RMSE scores for each training period day from September 16, 2020, to March 14, 2021, providing insights into how these models adapted and remained robust over time.

Subsequently, we turned our attention to evaluating these models' ability to predict performance on the mock exam using the same metrics: AUC, RMSE, and ACC. However, the mock exam posed a unique challenge as it featured entirely new questions that were not part of the training period. To address this challenge, we needed to estimate the difficulty of these mock exam questions. Our approach involved combining the entire training data set with a random selection of 60% of user data from the mock exam data set as the train set while designating the remaining 40% of user data from the mock exam data set as the test set. This allowed us to create a training set that encompassed all attempts, including those from the mock exam, for 60% of users. For the remaining 40% of

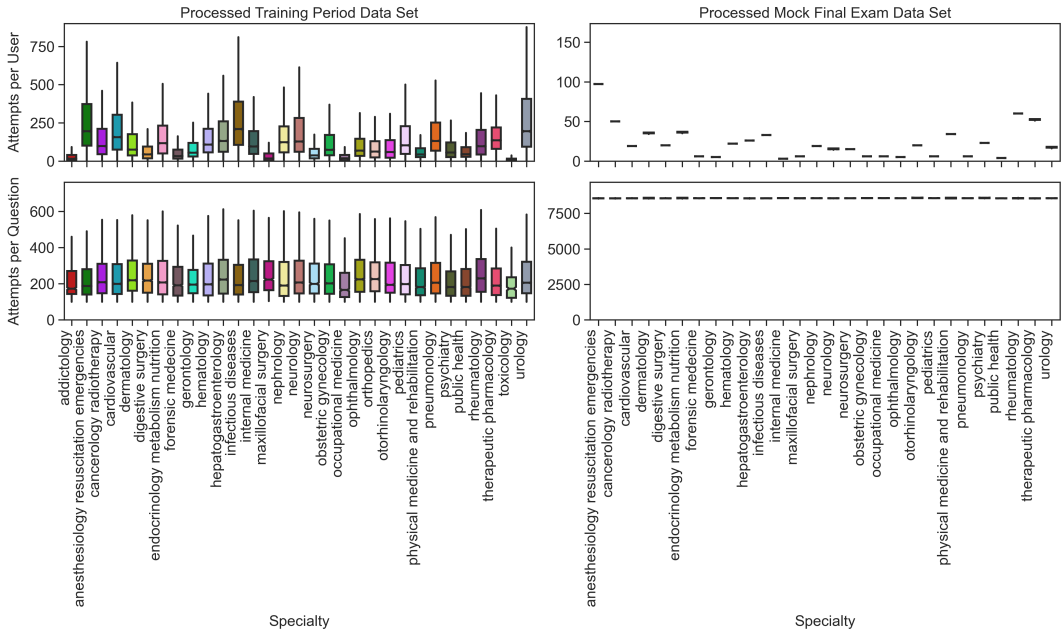

Figure 3: Number of Attempts by Each User and to Each Question across the 31 Medical Specialties.
The top left box plot shows the distributions of the number of attempts by each user across the 31 specialties. The bottom left box plot depicts the number of attempts to questions in each specialty. During the mock exam, all students took identical questions, resulting in quasi-uniform numbers of attempts given by students and received by questions (top and bottom right plots).

users, we included only their attempts from the training period in the train set. By doing so, when we ran the learner models on the training set, we obtained estimates of ability for all students and difficulty for all questions, which in turn enabled us to measure the models' prediction ability on the mock exam.

Following this data division process, the training subset comprised a substantial 24,152,933 entries, involving 8,616 unique users and 74,971 unique questions. In parallel, the test subset consisted of 1,268,752 entries, encompassing 3,447 unique users and 372 unique questions.

To assess the prediction ability of the models on the mock exam, after executing the models on the specified training set and obtaining difficulty estimates for all questions and ability estimates for all students, we evaluated its prediction performance on the test set. This evaluation capitalized on the stabilized estimates derived from the comprehensive training data set.

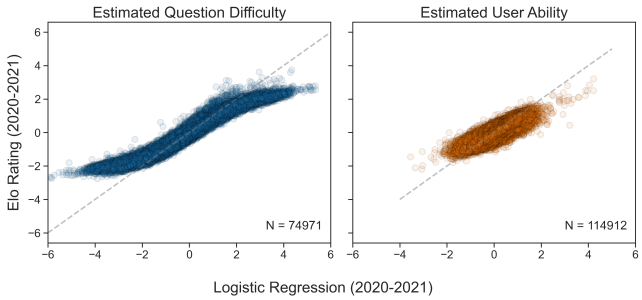

Figure 4: Comparing Logistic Regression and Elo Rating outcomes for question difficulty (left) and user ability across 31 specialties (right) in the same 2020-2021 dataset. Scatter plots illustrate the alignment, with $y = x$ lines for reference. The left plot displays Logistic Regression difficulty estimates on the x-axis and Elo Rating estimates on the y-axis. On the right, the plot contrasts user ability estimates, with Logistic Regression on the x-axis and Elo Rating on the y-axis. Sample sizes (N) are included in each plot.

3 RESULTS

3.1 Correlation between the Estimates

Figure 4 illustrates a notable positive correlation between the estimates of question difficulty ($r = 0.97$) and user ability ($r = 0.86$) derived from the Elo rating and logistic regression models for the same-year data. This strong positive correlation clearly indicates that both models converged toward similar final estimations regarding question difficulty and user ability.

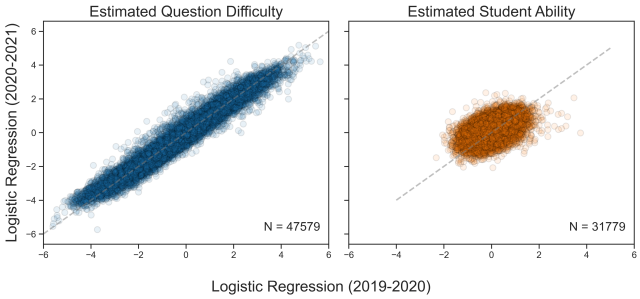

Figure 5: Comparing Logistic Regression Outcomes. Estimated question difficulty (left) and user ability on each of 31 specialties (right) in the two successive education years (2019-2020 and 2020-2021) using the Logistic Regression model. $y = x$ lines are given for reference.

Figure 5 shows the question difficulty and student ability estimations using the logistic regression in the two successive years. Specifically, for shared questions, we observe a robust positive correlation of 0.98, indicating that the difficulty levels of these questions remain relatively consistent over time. However, when it comes to students' abilities, the correlation, albeit positive at 0.54, is not as strong. This suggests that students' abilities in various specialties have undergone some changes over the years, as we anticipated.

3.2 Prediction Accuracy

Figure 6 presents a visual representation of the RMSE and AUC scores over 180 consecutive training days for each model. Findings indicate a substantial prediction accuracy advantage at the beginning of the training year when initializing the ability and difficulty values based on the data from the previous year. This advantage is reflected in an initial boost in average accuracy (+0.016 AUC) and a reduction in the average error (-0.008 RMSE) during the initial 30 days of training. However, it's worth noting that the initial disparity between the two model versions diminishes rapidly and becomes less than 1 point within a few days. By the end of the training period, the advantage of initializing with historical data becomes nearly negligible, with only a marginal improvement in average accuracy (+0.002 AUC) and a minimal reduction in average error (-0.002 RMSE) observed during the last 30 days of training.

Table 2 shows the prediction performance on the mock exam for the three models: Elo rating initialized at 0, Elo rating initialized historical data, and logistic regression. These results reveal that the three models show highly similar prediction accuracy, with a slightly better performance on the Elo rating model initiated with historical data over other models.

4 DISCUSSION

This research is an initial step in integrating the multi-concept Elo rating system into our medical training platform in order to achieve real-time estimates of user performance. The results demonstrated the Elo rating system's comparable prediction power to the logistic regression models, confirming its suitability for this specific data set.

The Elo rating system offers several significant advantages in the context of our medical training platform. Firstly, the multi-concept Elo rating system excels in estimating concept-level competencies, which is crucial for tailoring adaptive learning experiences in our data in which questions mostly require knowledge from multiple medical specialties. Secondly, it stands out as a computationally efficient and cost-effective option especially in real-time estimations, compared to logistic regression models which require processing a vast amount of previously collected data.

Given our primary objective of identifying the most effective prediction model for online applications in our platform, and considering the challenges associated with logistic regression in terms of online adaptability, our focus naturally shifted towards a more detailed comparison of the two versions of the Elo rating system: one starting from scratch at the beginning of the year, and one with difficulty and ability values initialized based on the previous year's data.

Our findings underscore that while the overall performance in predicting mock exam results remains very similar regardless of the initialization approach, a distinct advantage emerges for initialization based on historical data, particularly during the initial phases of iteration. This may be important in scenarios that demand accurate estimations from the outset, such as real-time or online applications. This approach of initializing the model with historical data enhances the model's ability to produce quicker and more precise estimates, thereby enhancing the reliability of personalized learning environments utilizing Elo rating systems.

4.1 Unique Characteristics of the Data set

Our data set has very broad knowledge components, made of 31 distinct medical specialties, each of which is a huge corpus of information. Moreover, this platform prioritizes comprehension over memorization: questions are hardly ever repeated twice. Students have to generalize their knowledge while attempting the questions. This departure from purely memory-based learning provides an excellent chance to assess the model's performance in dealing with non-repeated inputs. While this limited exposure to questions challenges standard prediction models used for repeated question iterations, it also allows us to evaluate the model's flexibility in settings where students rely on wider conceptual knowledge rather than memorized responses. Additionally, our platform allows students to personalize their own training experiences. They have the option of selecting the medical specialty in which they want to train and the type of questions, resulting in unique interaction patterns for each student. This adds another layer of complexity to our data set. In addition to these complexities, we also lacked control over learning occurring outside the platform.

Despite these challenges, it is remarkable that the Elo rating system has achieved significant prediction power for assessing the accuracy of students' future responses, with about 73.7% accuracy and 0.81 AUC. This demonstrates the model's adaptability and endurance in circumstances that deviate from standard educational data. In addition, the Elo rating system has shown good online prediction accuracy during training, right from the start when initializing with historical data, and after about 15-20 days when starting from scratch.

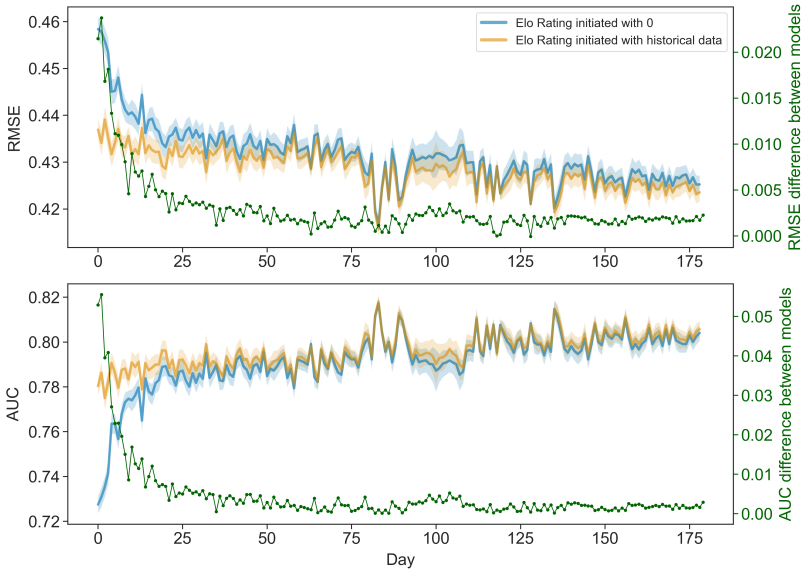

Figure 6: Comparing Models' Prediction Performance During Training.
RMSE (top) and AUC (bottom) values as a function of training days, for two versions of the Elo rating system. Shaded regions around the mean lines represent the 95% confidence intervals calculated using the standard error. A secondary y-axis on the right side illustrates the absolute difference between the two models with the green line plot.

Table 2: Comparing Models' Prediction Performance on Mock Exam

Model	RMSE (↓)	AUC (↑)	ACC (↑)	Precision (↑)	Recall (↑)	F1 (↑)
Elo rating initialized at 0	0.419	0.812	0.737	0.717	0.685	0.701
Elo rating initialized with historical data	0.418	0.813	0.738	0.720	0.684	0.702
Logistic regression	0.419	0.811	0.736	0.714	0.689	0.701

With the overarching goal of transforming our medical training platform into a personalized and adaptive format through knowledge tracing methods, we carefully considered the data set's characteristics. The Elo rating model stood out as a prime choice due to its simplicity, rapid parameter estimation, and real-time knowledge assessment capabilities. Its straightforwardness, coupled with widespread use in applications like online games and chess, makes it easily explainable compared to more complex models. An exemplar of transparency in a multivariate Elo version is evident in the study by [3]. The study demonstrates the feasibility of making the algorithm transparent to students, a practice that not only heightened their motivation to engage with the platform but also enhanced their trust in the recommendations provided. The implemented extensions further enhance adaptability to our data set's unique characteristics, offering optimization along with advantages in suitability and transparency. Notably, the multi-concept Elo rating model, in contrast to its single-concept version, acknowledging the non-transitive nature of skills, provides a realistic representation of learners' capabilities, crucial for accommodating interdependencies within medical specialties, potentially involving prerequisites. Thus, the multi-concept Elo rating model emerges as a fitting and transparent knowledge-tracing method for our complex data set.

4.2 Limitations and Further Work

One major limitation is that our tested models (logistic regression and Elo rating) do not consider the natural forgetting of knowledge over time, which is well-documented in human memory research dating back to Ebbinghaus in 1885 [13]. Incorporating learning and forgetting curves into prediction models, as shown in research like DAS3H [11] and MV-Glicko [2], which builds upon the multivariate Elo rating system [3], can improve the models' prediction accuracy.

Our study also suggests several promising future research directions. One area of focus is improving learning models to better assess question difficulty and student ability, especially for topics requiring knowledge of multiple concepts. This involves determining the importance of each concept in problem-solving and investigating how these concepts interact during the learning process. Additionally, comparing the Elo rating system with the Glicko rating model within our data set could provide insights into the role of learning and forgetting curves in student performance estimation.

Beyond the aforementioned research areas, a critical future direction involves implementing the Elo rating system for online recommendations regarding specialties and question difficulty. However, this endeavor presents challenges in platforms where questions often have multiple specialty tags. As underscored by the research

findings of [9], the premature revisiting of knowledge can exert detrimental effects on long-term memory. Consequently, during the recommendation phase, it becomes imperative to not only select questions aligned with the student's current needs but also to ensure that these questions do not encompass specialties that may not be relevant to the student's current stage of learning. To address this, we can consider a new approach that calculates students' abilities based on combinations of specialties, rather than individual ones. This new strategy could substantially enhance the model's efficacy when suggesting questions that align with students' learning requirements and are pertinent to their current learning stage. Such an approach would ensure that students are presented with a tailored set of questions that optimally support their progress while avoiding the unnecessary revisiting of topics that might hinder long-term retention—a crucial consideration for the success of an adaptive learning platform.

5 CONCLUSION

In conclusion, our study underscores the remarkable adaptability of the Elo rating system to the intricate challenges posed by a large, sparse, and multifaceted data set, where questions are tagged with multiple knowledge components. The Elo rating system, along with its enhanced version that leverages historical data for initial estimations, has exhibited a commendable level of prediction accuracy.

These results offer reassurance regarding the Elo system's robustness and versatility, emphasizing its capability to provide reasonable predictive value even in complex situations. This insight is crucial for the broader learning analytics community, providing confidence in the effectiveness of the Elo rating system as a predictive model in educational settings. Overall, the study contributes to the ongoing discourse on learning analytics methodologies, offering practical insights and encouraging further exploration of the Elo rating system's applicability in diverse learning scenarios.

ACKNOWLEDGMENTS

This research was funded by Agence Nationale de la Recherche, grants ANR-17-EURE-0017, ANR-10-IDEX-0001-02 and ANR-21-CE28-0030.

REFERENCES

[1] Ghodai Abdelrahman, Qing Wang, and Bernardo Nunes. 2023. Knowledge tracing: A survey. *Comput. Surveys* 55, 11 (2023), 1–37.
[2] Solmaz Abdi, Hassan Khosravi, and Shazia Sadiq. 2021. Modelling Learners in Adaptive Educational Systems: A Multivariate Glicko-Based Approach. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. 497–503.
[3] Solmaz Abdi, Hassan Khosravi, Shazia Sadiq, and Dragan Gasevic. 2019. A multivariate ELO-based learner model for adaptive educational systems. In *EDM 2019-Proceedings of the 12th International Conference on Educational Data Mining*. International Educational Data Mining Society, 228–233.
[4] Hamdan Alamri, Victoria Lowell, William Watson, and Sunnie Lee Watson. 2020. Using Personalized Learning as an Instructional Approach to Motivate Learners in Online Higher Education: Learner Self-Determination and Intrinsic Motivation. *Journal of Research on Technology in Education* 52, 3 (2020), 322–352.
[5] Shadi Aljawarneh and Juan A. Lara. 2021. Data Science for Analyzing and Improving Educational Processes. *Journal of Computing in Higher Education* 33 (2021), 545–550.
[6] Margit Antal. 2013. On the Use of Elo Rating for Adaptive Assessment. *Studia Universitatis Babes-Bolyai, Informatica* 58, 1 (2013), 29–41.
[7] Yigal Attali. 2014. A Ranking Method for Evaluating Constructed Responses. *Educational and Psychological Measurement* 74, 5 (2014), 795–808.
[8] Christopher Brooks and Craig Thompson. 2017. Predictive Modelling in Teaching and Learning. *Handbook of Learning Analytics* (2017), 61–68.
[9] Nicholas J. Cepeda, Edward Vul, Doug Rohrer, John T. Wixted, and Harold Pashler. 2008. Spacing Effects in Learning: A Temporal Ridgeline of Optimal Retention. *Psychological Science* 19, 11 (2008), 1095–1102.
[10] Wai-Lun Chan and Dit-Yan Yeung. 2021. Clickstream knowledge tracing: Modeling how students answer interactive online questions. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. 99–109.
[11] Benoît Choffin, Fabrice Popineau, Yolaine Bourda, and Jill-Jênn Vie. 2019. DAS3H: Modeling Student Learning and Forgetting for Optimally Scheduling Distributed Practice of Skills. In *Proceedings of the Twelfth International Conference on Educational Data Mining (EDM 2019)*. 29–38.
[12] Albert T. Corbett and John R. Anderson. 1994. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction* 4 (1994), 253–278.
[13] Hermann Ebbinghaus. 2013. Memory: A Contribution to Experimental Psychology. *Annals of Neurosciences* 20, 4 (2013), 155.
[14] Arpad E. Elo and Sam Sloan. 1978. *The Rating of Chessplayers: Past and Present*.
[15] Aritra Ghosh, Neil Heffernan, and Andrew S Lan. 2020. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2330–2339.
[16] G-H Gweon, Hee-Sun Lee, Chad Dorsey, Robert Tinker, William Finzer, and Daniel Damelin. 2015. Tracking student progress in a game-like learning environment with a monte carlo bayesian knowledge tracing model. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. 166–170.
[17] Sharon Klinkenberg, Marthe Straatemeier, and Han LJ van der Maas. 2011. Computer Adaptive Practice of Maths Ability Using a New Item Response Model for On-the-Fly Ability and Difficulty Estimation. *Computers & Education* 57, 2 (2011), 1813–1824.
[18] Jung Lee and Ok-Choon Park. 2008. Adaptive Instructional Systems. In *Handbook of Research on Educational Communications and Technology*. Routledge, 469–484.
[19] Lap-Kei Lee, Simon KS Cheung, and Lam-For Kwok. 2020. Learning Analytics: Current Trends and Innovative Practices. *Journal of Computers in Education* 7 (2020), 1–6.
[20] Robert V. Lindsey, Jeffery D. Shroyer, Harold Pashler, and Michael C. Mozer. 2014. Improving Students' Long-Term Knowledge Retention Through Personalized Review. *Psychological Science* 25, 3 (2014), 639–647.
[21] Wenting Ma, Olusola O. Adesope, John C. Nesbit, and Qing Liu. 2014. Intelligent Tutoring Systems and Learning Outcomes: A Meta-Analysis. *Journal of Educational Psychology* 106, 4 (2014), 901.
[22] Jan Papoušek and Radek Pelánek. 2015. Impact of Adaptive Educational System Behaviour on Student Motivation. In *Artificial Intelligence in Education: 17th International Conference, AIED 2015, Madrid, Spain, June 22-26, 2015. Proceedings* (Madrid, Spain), Vol. 17. Springer International Publishing.
[23] Jan Papoušek and Radek Pelánek. 2017. Should We Give Learners Control Over Item Difficulty?. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*. 299–303.
[24] Jan Papoušek, Radek Pelánek, and Vít Stanislav. 2014. Adaptive Practice of Facts in Domains with Varied Prior Knowledge. In *Educational Data Mining 2014*.
[25] Radek Pelánek. 2014. Application of Time Decay Functions and the Elo System in Student Modeling. In *Educational Data Mining 2014*.
[26] Radek Pelánek. 2016. Applications of the Elo Rating System in Adaptive Educational Systems. *Computers & Education* 98 (2016), 169–179.
[27] Chris Piech, Joel Bassen, Jennifer Huang, Surya Ganguli, Mehran Sahami, Leonidas J. Guibas, and Jascha Sohl-Dickstein. 2015. Deep Knowledge Tracing. In *Advances in Neural Information Processing Systems (NIPS)*. 505–513.
[28] Georg Rasch. 1960. *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests*.
[29] Sherry Ruan, Wei Wei, and James Landay. 2021. Variational deep knowledge tracing for language learning. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. 323–332.
[30] Dongmin Shin, Yugeun Shim, Hangyeol Yu, Seewoo Lee, Byungsoo Kim, and Youngduck Choi. 2021. Saint+: Integrating temporal features for ednet correctness prediction. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. 490–496.
[31] Behzad Tabibian, Utkarsh Upadhyay, Abir De, Ali Zarezade, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2019. Enhancing Human Learning via Spaced Repetition Optimization. *Proceedings of the National Academy of Sciences* 116, 10 (2019), 3988–3993.
[32] Wim J. van der Linden and Ronald K. Hambleton (Eds.). 2013. *Handbook of Modern Item Response Theory*. Springer Science & Business Media.
[33] Kurt VanLehn. 2011. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist* 46, 4 (2011), 197–221.
[34] Jill-Jênn Vie and Hisashi Kashima. 2019. Knowledge Tracing Machines: Factorization Machines for Knowledge Tracing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 750–757.
[35] Kelly Wauters, Piet Desmet, and Wim Van Den Noortgate. 2012. Item Difficulty Estimation: An Auspicious Collaboration Between Data and Judgment. *Computers & Education* 58, 4 (2012), 1183–1193.