Journal of Computer Assisted Learning

WILEY

Journal of Computer Assisted Learning

ORIGINAL ARTICLE

Investigating the Influence of Training Difficulty on the Learning Outcomes of Medical Students

Erva Nihan Kandemir[1] | Jill-Jênn Vie[2] | Adam Sanchez-Ayte[3] | Olivier Palombi[4] | Franck Ramus[5]

[1]Laboratoire De Sciences Cognitives et Psycholinguistique, Département D'etudes Cognitives, Ecole Normale Supérieure, CNRS, PSL University, Paris, France | [2]Soda Team, Palaiseau, France | [3]Université Numerique en Santé et Sport (UNESS), Inria, LIG, France | [4]Université Grenoble Alpes, Grenoble, INP, CNRS, Saint-Martin-d'Hères, France | [5]Laboratoire De Sciences Cognitives et Psycholinguistique, Département D'etudes Cognitives, Ecole Normale Supérieure, EHESS, CNRS, PSL University, Paris, France

Correspondence: Erva Nihan Kandemir (erva.nihan.kandemir@ens.psl.eu)

Received: 10 August 2024 | **Revised:** 7 April 2025 | **Accepted:** 22 November 2025

Keywords: adaptive learning | computer-assisted learning | learning analytics | learning strategies | optimal training difficulty

ABSTRACT

Background: Determining an optimal training difficulty level for the best learning outcome is a crucial goal for adaptive educational systems. The literature supports the Inverted *U*-shape Hypothesis, suggesting that the ideal challenge level for learning is neither too easy nor too difficult. However, this optimal point depends on the type of training and response modality and may vary across domains, necessitating thorough examination before implementing adaptive learning procedures.

Objectives: This study aimed to investigate the influence of training difficulty on the learning outcomes of French medical students.

Methods: Using data from a national educational platform, we explored the influence of the mean question difficulty encountered during training, relative to individual student ability, on the learning outcomes of medical students across diverse medical specialties. Importantly, the mean difficulty level varied randomly between students on this platform, mirroring a quasi-experimental design and enabling a thorough exploration of these effects. We first employed the Elo rating system to estimate the difficulty of platform questions and the evolution of students' abilities. A linear mixed-effects model was then used, with final exam performance as the main outcome and mean relative question difficulty during training (linear and quadratic terms) as the main predictor.

Results and Conclusions: Results showed a significant negative quadratic effect of mean relative difficulty on the final exam performance, revealing optimal difficulty levels for each medical specialty. Additionally, the analysis demonstrated that students with high abilities displayed a more pronounced inverted *U*-shaped relationship between training difficulty and final exam scores. This study advances our understanding of optimal training difficulty in the complex realm of medical education, by emphasising the need to acknowledge variability across medical specialties and student abilities.

1 | Introduction

The pursuit of knowledge, from initial acquisition to continuous practice, requires careful consideration of various factors influencing learning. Previous research has consistently highlighted the pivotal role of learners' retrieval practice behaviour in predicting learning outcomes (Hattie 2012; Roediger III and Karpicke 2006; Sinha et al. 2014; De Morais et al. 2014; Schütt et al. 2023). Thus, researchers and educators are keen on identifying conditions that promote optimal training and, consequently, optimal learning. Among multiple factors such as learner motivation, prior knowledge, feedback, learning schedule, learning

© 2025 John Wiley & Sons Ltd.

Lay Summary

What is currently known about this topic:

- The concept of an optimal level of challenge in learning, known as the Inverted U-shape Hypothesis, suggests that the best learning outcomes occur when tasks are neither too easy nor too hard.
- The exact level of optimal difficulty can vary based on the subject and the individual's abilities

What this paper adds:

- Explores the impact of training difficulty on learning outcomes specifically for French medical students using a national educational platform.
- Confirms the Inverted *U*-shape Hypothesis in French medical education.
- Reveals that optimal training difficulty varies across medical specialties and individual differences.

Implications for practice and/or policy:

- Supports personalised difficulty adjustments in educational systems to enhance learning outcomes.
- Recommends adaptive learning technologies that tailor the difficulty to student abilities and subject complexity.

tasks, and so forth, training difficulty emerges as particularly crucial.

Searching for the ideal balance between challenge and achievability in training has been a central theme in educational research for many years. The Zone of Proximal Development idea of Vygotsky (1978) laid the foundation for this ongoing investigation. The theory proposed that the best learning occurs at the edge of the learner's competence, and research has shown that teaching skills within this zone leads to better performance compared to those outside it (Zou et al. 2019). This zone encourages learners to cope with new concepts while remaining achievable enough to prevent discouragement and frustration. Yet, given that each learner possesses different skills and levels of prior knowledge of the learning material, Vygotsky's theory inherently implies that the term "optimal difficulty" is a value that depends on the learner's ability. Thus, in practice, its original definition provides little guidance to determine an adequate difficulty level.

In an attempt to provide more concrete guidelines to achieve this balance and operationalise the concept of optimal difficulty, Rosenshine (2012) famously claimed that an 80% success rate in guided practice is optimal for learning. However, this claim was predominantly grounded in limited and small-scale evidence. It originated from observational "process-product" studies in elementary school reading and mathematics (Good and Grouws 1977; Gersten et al. 1982). Consequently, the reliability and generalizability of these conclusions across diverse educational contexts remain ambiguous and under-explored.

Due to the inherent variability in learners' competence, traditional approaches to determining the appropriate level of challenge in learning materials often relied on handcrafted rules devised by domain experts. However, the advent of Adaptive Learning Systems (ALS) (Park and Lee 2003) has revolutionised this methodology by employing adaptive algorithms and extensive question banks to provide dynamic and personalised learning experiences. There are several examples of adaptive online learning systems for example, for learning factual knowledge in the field of geography (Pelánek et al. 2017) or mathematics (Klinkenberg et al. 2011). Research consistently highlights the superiority of ALS over non-adaptive counterparts, demonstrating markedly improved learning outcomes (Papoušek et al. 2016a; Ma et al. 2014).

In particular, adjustment of the difficulty level in training exercises to accommodate diverse prior knowledge and learning rates of learners holds significant promise. Previous research has suggested that selecting the appropriate difficulty level for practice exercises yields positive effects on how students perceive their own ability (Power 2019), learning gains (Sampayo-Vargas et al. 2013) and learning experiences (Kostons et al. 2010; Deterding and Cutting 2023). Consequently, various methods have been explored to adjust difficulty levels according to the user's needs. One approach involves granting users autonomy to select their next difficulty level—a process known as self-determined adjustment (Westlin et al. 2019). Alternatively, another method involves estimating the user's ability level based on their past performance and automatically presenting exercises tailored to that level. For example, presenting more challenging exercises to more advanced or high-performing students and simpler exercises to beginners or those who struggle. This process is often referred to as Dynamic Difficulty Adjustment (DDA). Research has underscored the benefits of both integrating learner choice into learning activities (Salden et al. 2006) and implementing dynamic adjustment strategies (Romero et al. 2006; Sampayo-Vargas et al. 2013; Zhang and Goh 2021), resulting in enhanced learning outcomes across diverse domains. However, a recent study (Schütt et al. 2024) comparing dynamic difficulty adjustment with self-determined difficulty levels in open-ended learning tasks found no significant differences in learning gains. Additionally, a separate study (Papoušek and Pelánek 2017) suggested that instead of providing learners with the option to adjust difficulty, developing methods for automatic adjustment of target difficulty may offer a more effective solution.

Despite the evident advantages of adaptive learning systems that customize difficulty levels based on individual learner abilities, a fundamental question persists: what defines an appropriate level of difficulty to target? In many adaptive learning systems, these target difficulty levels are externally set by developers based on various considerations.

One example of this approach is the staircase method, which adjusts task difficulty to maintain a fixed error rate during learning, typically targeting an accuracy range of 80%–85% (García-Pérez 1998). This choice, however, is primarily based on intuition rather than empirical validation. Furthermore, enforcing a single target difficulty level applicable to all learners, content, and contexts may not be optimal. At the very least, the optimal level should depend on the chance level of correct responses, which is influenced by the question/

response format. Additionally, it may be influenced by domain complexity, student skill, and other context-dependent factors. For instance, an 80% success rate may not represent the same level of difficulty across different learning activities, such as problem-solving tasks, generative activities with open-ended responses, and multiple-choice questions with varying numbers of propositions.

These complexities underscore the critical need for a more rigorous investigation into the validity and effectiveness of predetermined challenge levels. Therefore, before integrating an adaptive algorithm into learning systems, a thorough exploration of the optimal training difficulty becomes indispensable for maximizing learning outcomes in a given learning task and platform.

Recent studies exploring optimal training difficulty have provided support for the Inverted-*U* shape Hypothesis. A large-scale observational study (Abuhamdeh and Csikszentmihalyi 2012) involved thousands of online chess players. The results revealed that players derived the greatest enjoyment when they competed against opponents who were slightly more skilled than they were (with a winning probability of approximately 20%), reflecting a correlation with the Inverted-*U* pattern. One notable limitation of this study was that players were not randomly assigned opponents; instead, they had the autonomy to select opponents based on their chess rank, allowing them to control their expected game difficulty. Furthermore, the study focused on an intrinsically motivated activity, and the outcome of interest was the engagement of online chess players. It therefore says nothing about what they learned. In contrast, school and work-related activities often lack intrinsic motivation, being driven by an obligation to learn. Thus it is important to bear in mind that the optimal difficulty for enjoyment and motivation may not be the same as those for learning. Notably, Cao et al. (2022) demonstrated that motivation plays a mediating role in the relationship between difficulty and learning in educational gamification. Nonetheless, the optimal conditions for activities with intrinsic motivation may not be the same as those for less intrinsically motivated activities.

In educational contexts, evidence suggests that the impact of difficulty on learning and engagement can be conflicting. In the context of their simple educational game, Lomas et al. (2013) found that simpler problems resulted in higher engagement but lower levels of learning. Similarly, research conducted using the Math Garden software Jansen et al. (2013) compared three conditions with target success rates of 60%, 75%, and 90%, revealing that the easiest condition resulted in the most effective learning (mediated by the number of solved tasks). Another recent study (Papoušek et al. 2016b) examining question difficulty in the learning of declarative knowledge arrived at a different conclusion, suggesting that more challenging questions enhance learning and long-term engagement, while easier questions are better for short-term engagement. Thus, while there is a general understanding that tasks are most engaging when they strike a balance between being too easy and too hard, the precise optimal point and the specific impact of a one-point increase in training difficulty on the learning outcome remain open questions.

Recently, the concept of optimal difficulty was extended beyond human learning in the domain of machine learning. Wilson et al. (2019) provided evidence that an 85% training accuracy served as a "sweet spot" for optimal learning in binary classification tasks. This theoretical result aligns with the inverted U-shape hypothesis, and supports the idea that relatively easy tasks promote better learning, consistent with Jansen et al. (2013) but not with Lomas et al. (2013) and Papoušek et al. (2016b).

Collectively, these studies emphasise a key principle: optimal learning occurs when difficulty is calibrated appropriately. Nevertheless, diverse findings suggest that a universal optimal difficulty may not exist for all individuals and across every domain.

Therefore, this study aimed to provide a robust answer to the question of optimal training difficulty, within the specific context of multiple-choice questions in a complex domain: medical education. It relies on an online non-adaptive training platform that encompasses an extensive knowledge corpus, significant overlap between knowledge domains, and a very large bank of multiple-choice questions. We specifically addressed the following two research questions (RQs):

RQ1. Can an optimal level of training difficulty be determined for medical students using our online platform?

RQ2. If so, what is it, and does it vary between medical specialties?

To address our research questions, we applied the Elo rating system (Elo 1978) to our dataset in order to estimate both the proficiency levels of students as they trained and the difficulty levels of questions within the platform. We then capitalised on the fact that on this platform, training questions are drawn at random. Therefore, students are exposed to questions of random difficulty. Over a training semester, by pure chance, some students will be exposed to more difficult questions than others, on average. Therefore, the mean difficulty level of the questions on which students trained can be considered an instrumental variable in a quasi-experimental design. Thus, we conducted statistical analyses to examine the impact of the mean training difficulty (relative to students' competence) on the final exam performance, aiming to identify the optimal difficulty level that leads to the best learning outcome, as measured on a final exam.

We hypothesized that the relation between mean training difficulty and exam performance is non-linear, with a maximum at an optimal difficulty level. Moreover, recognising the inherent diversity in difficulty levels across medical specialties, we further hypothesized that the optimal difficulty level might differ to some extent between specialties. Of course, we also expect the main effects of student ability and training intensity on exam performance. Since these factors may affect both

training and final performance, they will be included in our analyses as covariates.

2 | Method

2.1 | Study Design and Platform Features

This study adopted a quasi-experimental research design, utilising data sourced from the Banque Nationale d'Entraînement (BNE) digital learning platform. The subsequent section provides a comprehensive description of the BNE platform, its operational mechanisms, and the specific features of the question bank. A detailed presentation of the dataset is given in the following section.

2.1.1 | BNE Digital Learning Platform

The Banque Nationale d'Entrainement (BNE) digital learning system is an online platform managed by Université Numérique en Santé et en Sport (uness.fr) that is widely used by more than 8000 medical students from all French universities in each academic year from the 2nd to the 6th year. It is used for the administration of exams as well as for student training. The training on this platform is entirely student-driven and involves tackling multiple-choice questions across 31 different medical specialties, which helps students test and reinforce their knowledge in specific areas of medicine and prepare for the ECN (Épreuves Classantes Nationales) national final exam. This high-stakes exam, which takes place at the end of the sixth academic year, is the main determinant of students' access to residency in their chosen specialty.

The platform also incorporates a feedback system designed to enhance the learning experience for students. After each test, students receive corrective feedback on the correctness of their answers and, in a subset of questions, detailed explanations of why some answers are right or wrong.

Concretely, students who initiate a training session on the BNE digital learning platform freely select (1) the type of questions (isolated questions, progressive medical files, critical article reading, see description below); (2) the medical specialties they want to train on; (3) the number of questions of each type in each specialty. They can generate multiple training sets and train at their own pace as long as they want. This flexibility enables them to tailor their training program to their individual preferences and needs.

Importantly, although students are free to select the desired specialties and test types for training, the specific questions of each type and specialty are drawn at random from the question bank. Thus, students may be exposed by chance to questions of higher or lower difficulty on average. The mean difficulty level of the questions on which students trained can therefore be considered as an instrumental variable in a quasi-experimental design. We exploit this random allocation to test the effect of mean question difficulty during training (relative to students' current ability) on final exam outcomes.

2.1.2 | Question Bank

Past exam questions, along with those designed for training purposes, are incorporated into a comprehensive question bank, constituting an extensive collection of approximately 2,313,023 multiple-choice questions across 31 medical specialties. This huge question bank is easily accessible to students within the platform, offering a valuable and rich training resource to prepare for the ECN. The training tests drawn from the question bank on the platform are categorised into three main types:

1. **Isolated Questions Test (IQ)**: Classic multiple-choice questions.
2. **Progressive File Test (PF)**: A set of approximately 15 interconnected multiple-choice questions, where each question reveals new medical information about a patient, thus simulating a realistic medical situation.
3. **Critical Article Reading Test (CAR)**: A set of questions based on an article.

Furthermore, within each of these test categories, questions may be either SAQs (Single Answer Questions) or MAQs (Multiple Answer Questions).

It is also worth highlighting that the average number of specialties linked to each question is 1.58. This underscores that a considerable portion of questions (34.5%) in the question bank are tagged with multiple medical specialties, implying that answering them accurately requires knowledge from several medical specialties. Another notable aspect of the question bank is that, while it comprises questions with varying numbers of propositions, a significant majority of the questions—98.2% of training questions and 97.3% of ECN questions—consist of five propositions.

2.2 | Dataset

We extracted detailed records of student interactions from the BNE, including every attempt made by students to answer questions. The data included anonymized student IDs, question IDs, specialties tagged by the question, timestamp of the attempt, number of propositions in the question, test type, and score obtained in each attempt. This approach ensures the privacy and confidentiality of all students involved.

To test our hypothesis, we focused our analysis on the educational year 2020–2021, which was the most recent accessible dataset at the time of our analysis. Although direct access to the official ECN national final exam via the BNE platform was not available, we had access to a final exam known as ECNp (ECN préparatoire). This mock exam, typically held in mid-March (specifically on March 15th, 16th, and 17th, 2021, for the 2020–2021 academic year), closely mirrors the format of the actual ECN national final exam, and therefore served as an external outcome identical for all 6th-year students. We therefore focused on 6th-year students, and we will now refer to this mock exam as the final exam.

TABLE 1 | Data set summary.

Data	Period	Total Nb Attempts	Nb Students	Nb Questions	Nb Specialties	Sparsity (Student, Question)
Training Period Dataset	16.11.2020–14.03.2021	18,503,110	8633	323,701	31	0.99
Final Exam (ECNp) Dataset	15–17.03.2021	3,178,527	8636	372	28	0.01

In order to obtain a dataset allowing for a robust estimation of question difficulty and student ability in each specialty, and allowing for a robust analysis of the relation between training conditions and exam results, we applied a number of steps of data cleaning and selection, as described below.

2.2.1 | Data Cleaning

After obtaining the raw data for the 6th-year students for the 2020–2021 educational year, we first conducted a series of preprocessing steps on this dataset. These steps were executed in the following sequence:

1. Removal of duplicated rows: Only 144 duplicated rows out of 27,557,031 were removed.
2. Exclusion of questions without any tagged medical specialty: 0.28% of the questions (making 0.21% of attempts) lacked specialty tags.
3. Binarisation of question grading: BNE employs a sophisticated grading scheme depending on the number of correct and incorrect answers ticked. We binarised scores by mapping all scores other than 1 to 0.

2.2.2 | Training Period Dataset

Table 1 provides an overview of the key characteristics of our training period dataset.

An essential variable highlighted in Table 1 is the student-question sparsity, indicating the proportion of missing values in the student-question interaction matrix. The table illustrates that the dataset is extensive but shows significant sparsity ($Sparsity\ (student, question) = 0.99$) during the training period. This reflects that there are considerably more available questions than any individual student can attempt, and different students most often attempt different randomly selected questions.

2.2.3 | ECNp—Final Exam Dataset

Table 1 also shows characteristics of the final exam (ECNp) covering 28 different medical specialties (addictology, orthopaedics, and toxicology were not tested in the final exam). A total of 3,178,527 user interactions were recorded, with a sparsity value of 0.01, suggesting that nearly all students attempted almost every question in the exam.

It is important to emphasise that the final exam is not a retest of training questions, but rather a brand new set of questions designed to test students' ability to generalise their medical knowledge acquired in training to unfamiliar questions. This reflects the actual ECN exam's focus on broader understanding over rote memorization. This differs from many other experimental studies where learning is assessed mainly on repeated questions.

It is also noteworthy that the distribution of questions across medical specialties in the final exam is not uniform. Figure 1 illustrates the varying number of questions allocated to each specialty in the final exam, along with the corresponding proportion of correct answers.

2.2.4 | Data-Filtering for Statistical Analysis

Although the Elo rating system was initially applied to unfiltered data to obtain the mean relative difficulty during the 4-month training period for the total of 253,036 student-specialty pairs, for the statistical analysis we implemented filtering procedures to enhance the reliability of results.

First, to ensure a robust measure of learning outcome (final exam score), the analysis was limited to medical specialties with a sufficient number of questions (≥ 20) in the final exam, resulting in 13 out of 31 specialties being included (see Figure 1).

Note that all students did not necessarily train on all specialties and did not always take enough questions in each specialty to allow for a reliable estimation of their ability in all specialties. Therefore, each student's training can only be analysed on a specific subset of medical specialties. Student-specialty pairs are thus the unit of our statistical analysis. Since our analysis focuses on within-subject differences between medical specialties, we excluded students who had not undergone training in at least two of the 13 selected medical specialties, with a minimum of 100 attempts each, to ensure a reliable estimation of online student ability from the Elo rating system. This led to the removal of 15% of students from the analysis.

As a result of these multiple data cleaning and selection steps, the final dataset available for analysis included 6451 students, resulting in 45,436 student-specialty pairs and a 46% sparsity of the student-specialty matrix.

2.3 | Students' Ability and Question Difficulty Estimations

Traditionally, models for students' ability and question difficulty estimations trace back to Rasch's single-parameter logistic

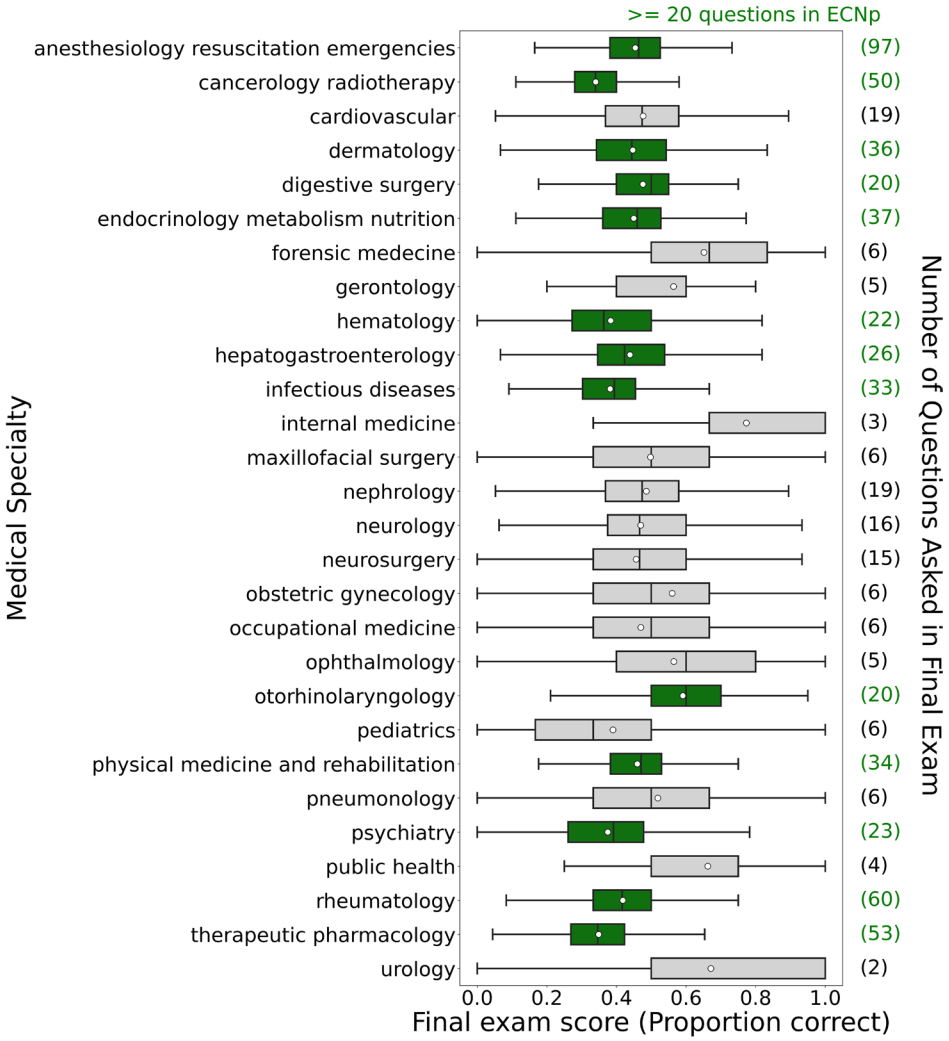


FIGURE 1 | Horizontal box plot illustrates the variability in correct answer rates among different medical specialties within the ECNp final exam. The parentheses on the right y-axis denote the number of questions posed in each specialty within the ECNp exam. Only specialties with a sufficient number of questions (≥ 20), shown in green, are included in the analysis.

regression formulation (Rasch 1960). Item Response Theory (IRT), based on this logistic model and its variations (van der Linden and Hambleton 2013), has been extensively used in educational applications, providing robust estimations (for a recent overview, see Wauters et al. (2010)).

While IRT is a well-established method, alternative approaches for estimating task difficulty and student ability have emerged. These include Markov process models (Gweon et al. 2015), deep knowledge tracing models (Chan and Yeung 2021; Piech et al. 2015), and rating systems (Pelánek 2016), each documented in existing literature.

In our context, because we were analysing training data over several months, it was important to have an up-to-date estimate of each student's ability each time they answered a new question, hence the need for dynamic estimations. This led us to focus on rating systems, and more particularly on the Elo rating system, as described by Pelánek (2016). The Elo rating system stands out as an algorithmically and computationally more economical approach compared to other models that necessitate processing vast amounts of previously collected data. Moreover, it has found widespread application in educational technologies (Klinkenberg et al. 2011; Pelánek 2016; Attali 2014) and has been validated as a suitable online tool without the need for extensive question pretesting on large sample sizes (Papoušek et al. 2014; Antal 2013; Kandemir et al. 2024).

2.3.1 | Elo Rating System for Question Difficulty and Online Student Ability Estimations

Initially designed for ranking chess players, the Elo rating system (Elo 1978) has found a new application in educational settings. In this context, students and learning materials (e.g., questions) are treated as opponents, each assigned ratings that reflect their abilities and difficulties, respectively. This repurposing allows the system to predict ratings for students and questions, functioning as an assessment tool for both student ability and question difficulty in educational environments.

In its standard educational formulation, every student s has an ability parameter θ_s and each question q has a difficulty

parameter θ_q. The probability of student s successfully answering a multiple-choice question q, denoted as $\Pr\left(a_{sq}=1|\,\theta_s,\theta_q\right)$, can be obtained using a logistic function σ of the discrepancy between student ability and question difficulty:

$$\Pr\left(a_{sq}=1|\,\theta_s,\theta_q\right)=\sigma\left(\theta_s-\theta_q\right)=\frac{1}{1+e^{-\left(\theta_s-\theta_q\right)}} \quad (1)$$

The system then dynamically adjusts the student's ability and the question's difficulty parameters after each interaction based on the disparity between the estimated probability correct ($Pr\left(a_{sq}=1\right)$) and the actual score (a_{sq}).

These updates are defined by the following formulas for question difficulty (θ_q) and for student ability (θ_s):

$$\begin{aligned}\theta_q &:= \theta_q + K\left(\Pr\left(a_{sq}=1|\,\theta_s,\theta_q\right)-a_{sq}\right)\\ \theta_s &:= \theta_s + K\left(a_{sq}-\Pr\left(a_{sq}=1|\,\theta_s,\theta_q\right)\right)\end{aligned} \quad (2)$$

where K is a constant value that determines the degree of update sensitivity based on the student's most recent attempt.

Thus the Elo rating system continuously adjusts student and question parameters through ongoing interactions. This system provides both current estimations of difficulty and ability parameters after each interaction, as well as final versions of these parameters at the end of the simulation. In this paper, we refer to the dynamically estimated ratings after each attempt as online difficulty and online student ability, and to the ratings stabilised at the end of the training as final student ability and final question difficulty.

Following the establishment of the standard Elo rating system formulation, the system has undergone various extensions (Pelánek 2016; Abdi et al. 2021). In a prior investigation, the multivariate version of the Elo rating system was adapted to suit the current database (Kandemir et al. 2024). This version accommodates situations where learning items can be tagged with multiple knowledge components, enabling a more nuanced estimation of student abilities. Similar to the standard Elo rating system, the predictive efficacy of the multivariate version has been established in previous studies (Abdi et al. 2021; Kandemir et al. 2024). These studies, drawing upon data from varied learning platforms, underscore its reliability in accurately estimating both student and question-related parameters, thereby highlighting its robustness.

As mentioned before, in the BNE, one question may be tagged by multiple specialties. Thus, to effectively account for student abilities within each of these tagged specialties, we employed the multi-concept extended version of the Elo rating system, as introduced by Abdi et al. (2021). The estimation of a student's online ability on a question involved averaging the student's online abilities across all specialties linked to the question. The details and exact formulation of the multivariate Elo rating system adapted to our dataset have been described in a previous work (Kandemir et al. 2024). Therefore, in this current study, we will utilise it as a "black box" model, focusing on its applications and outputs rather than its internal workings. The predictive efficacy of this multivariate Elo model within our dataset has also been evaluated in the aforementioned previous study. This was achieved by training the model on training period data and subsequently utilising it to predict the outcomes of mock final examinations. The high predictive performance (AUC = 0.81, RMSE = 0.42) aligns with findings from previous research (Abdi et al. 2021; Klinkenberg et al. 2011; Wauters et al. 2011). Here, the Area Under the Curve (AUC) measures the model's ability to distinguish between classes, while the Root Mean Square Error (RMSE) reflects the average magnitude of prediction errors, with smaller values indicating better predictions.

2.4 | Measures

Following the derivation of estimations for both students' abilities and question difficulties using the Elo rating system on the training data, we computed the following measures for each student–specialty pair:

Mean Relative Difficulty (Main predictor): In order to measure how challenging a student's training period was, we calculated the relative difficulty of each attempt by subtracting the student's estimated online ability on the attempt (averaged across all specialties tagging this particular question) from the estimated final difficulty of the question in the training period dataset. Whereas it is important to take online student ability at each attempt into account because it is expected to improve with practice, we used final rather than online question difficulty because it is presumed to be a stable property whose estimation is most reliable after many attempts. We then averaged the relative difficulty for each attempt that the student made on questions from a given specialty over the entire training period, yielding a mean relative difficulty for each student-specialty pair.

For a student S who attempted n questions in a given specialty k:

$$\text{Mean Relative Difficulty}_{S,k}=\frac{1}{n}\sum_{i=1}^{n}\left(\delta_i-\theta_{s,k,i}\right)$$

where:

- δ_i: The final difficulty estimation of question i, considered stable due to many attempts
- $\theta_{s,k,i}$: The online ability of student S in specialty k, which is required to solve question i, at the time of the attempt

For example, if a student with an online ability of 1 (positive values representing higher ability) took a question with a difficulty rating of 2 (positive values representing more difficult questionss), the relative difficulty of this action would be $(2-1=1)$. A positive mean relative difficulty score indicates that a student was exposed to relatively challenging questions during their training period for the given specialty, while a negative score suggests a relatively easier training period.

Final Exam Score (Main outcome): The learning outcome was the proportion of correct responses obtained by each student in each specialty in the final exam. It is calculated by using the Final Exam (ECNp) Dataset.

Given the observational nature of the study design and the reliance on Elo rating system estimates for both student ability and question difficulty, controlling for all potential dependencies—such as students' interactions with the platform and their inherent abilities—presented a challenge. One notable concern was the potential correlation between the mean relative difficulty and the final student ability. It was anticipated that higher-ability students would, on average, encounter lower mean relative difficulty during training. Consequently, a positive association between mean relative difficulty and exam score was likely, as higher-ability students tend to achieve higher scores in the final exam (Hattie 2012). In order to control for this confounding factor, our statistical model was adjusted for final student ability in each specialty. Additionally, we accounted for training intensity, measured by the number of questions attempted per specialty during training, as prior research has demonstrated that student engagement with practice questions significantly impacts learning outcomes (Lin et al. 2019).

Final Student Ability (Covariate): The final student ability in each medical specialty was the last estimation of their ability at the end of the training period. It is calculated by using the training period dataset.

Number of questions taken by specialty (Covariate): To account for the impact of training intensity on final exam outcomes, we included the number of questions taken by each student in each specialty and its interaction with final student ability as covariates in our model. This allowed us to better take into account the combined effects of training intensity and final student ability on exam performance. It is calculated by using the training period dataset.

After calculating these measures, we combined them for each student-specialty pair. Subsequently, a data-filtering process (refer to Section 2.2.4) was applied to refine this combined dataset. The resulting filtered data was then used for the statistical analysis.

2.5 | Analyses

2.5.1 | Statistical Model

We employed a linear mixed-effects model to investigate the influence of mean relative question difficulty on final exam scores. This analysis was conducted using the lme4 package (Bates et al. 2015) (Version 1.1–31) in R (Version 4.2.2), and the restricted maximum likelihood (REML) estimation method to account for the hierarchical nature of the data. Additional packages used for data wrangling, visualisation, and diagnostics included dplyr (Wickham et al. 2023), ggplot2 (Wickham 2016), performance (Lüdecke et al. 2021), sjPlot (Lüdecke 2021), and lmerTest (Kuznetsova et al. 2020). A complete list of all packages used, along with their versions, is provided in the openly shared repository.

The Final Exam Score was regressed on both linear and quadratic Mean relative difficulty as fixed effects, to test the hypothesis of a non-linear effect with a maximum.

Given the observational nature of the study design and the reliance on Elo rating system estimates for both student ability and question difficulty, controlling for all potential dependencies posed a challenge. One notable concern was the potential correlation between mean relative difficulty and student ability. Since relative difficulty is the difference between question difficulty and student online ability, it was anticipated that students with higher ability would, on average, end up with lower mean relative difficulty during training. Consequently, a negative association between mean relative difficulty and exam scores was expected, as students with higher final ability would tend to achieve higher scores in the final exam. In order to control for this confounding factor, our statistical model was adjusted for final student ability in each specialty.

An interaction between final student ability and both the linear and quadratic terms of mean relative difficulty was included to explore whether the impact of mean relative difficulty varied with final student ability.

The model also included the number of questions taken by each student in each medical specialty as a covariate. Moreover, we included an interaction between final student ability and the number of questions taken by students in each medical specialty, in order to test the hypothesis that training intensity may have a different effect depending on final student ability. To facilitate the interpretation of interaction terms, the number of questions taken in each specialty was mean-centred, while other predictors in the model were already centred around a mean of approximately zero (see Table 2).

We addressed the within-student between-specialty design and the complex structure of our dataset by further including:

Random Intercepts for Students. Capturing inherent variability in student performance across the dataset, independent of the specialty-specific influences.

TABLE 2 | Descriptive statistics.

Filtered data used in the model		
Number of students		6451
Number of medical specialties		13
Number of student-specialty pairs		45,436
Sparsity in the student-specialty matrix		46%
Descriptive statistics of the model variables		
	M	**SD**
1. Mean relative difficulty	−0.09	0.43
2. Final student ability	0.07	0.41
3. Number of questions taken by specialty	280	236
4. Final exam score (proportion correct)	0.45	0.13

TABLE 3 | Multicollinearity check results.

Term	VIF	VIF 95% CI	Increased SE	Tolerance	Tolerance 95% CI
Mean relative difficulty	1.27	[1.25, 1.28]	1.13	0.79	[0.78, 0.80]
Number of questions taken by specialty	1.32	[1.31, 1.34]	1.15	0.76	[0.75, 0.76]
Mean relative difficulty^2	3.80	[3.74, 3.86]	1.95	0.26	[0.26, 0.27]
Final student ability	1.23	[1.22, 1.24]	1.11	0.81	[0.80, 0.82]
Mean relative difficulty: Final student ability	3.83	[3.77, 3.89]	1.96	0.26	[0.26, 0.27]
Mean relative difficulty^2: Final student ability	1.17	[1.15, 1.18]	1.08	0.86	[0.85, 0.87]
Number of questions taken by specialty: Final student ability	1.37	[1.35, 1.39]	1.17	0.73	[0.72, 0.74]

Random Intercepts and Slopes for linear and quadratic terms of Mean Relative Difficulty. This allowed the influence of mean question difficulty on exam scores to vary across specialties.

Our linear mixed-effects model was thus formulated as follows (in R syntax):

$$\begin{aligned}&\text{final_exam_score} \sim \\ &\text{mean_relative_difficulty}^{*}\text{final_student_ability} + \\ &\text{mean_relative_difficulty}^{2*}\text{final_student_ability} + \\ &\text{nb_questions_taken_by_specialty}^{*}\text{final_student_ability} + \\ &(1|\,\text{student}) + \\ &(1 + \text{mean_relative_difficulty} + \text{mean_relative_difficulty}^2|\,\text{specialty})\end{aligned}$$

3 | Results

3.1 | Descriptive Statistics

Table 2 displays the descriptive statistics of the variables under analysis.

We assessed potential multicollinearity among all predictor variables using the Variance Inflation Factor (VIF). A VIF value exceeding the commonly accepted threshold of 5 indicates potential multicollinearity issues (Kim 2019; Sheather 2009). The results, presented in Table 3, show that all predictor variables included in our model have VIF values between 1.17 and 3.83. These values suggest that our model does not suffer from significant multicollinearity concerns.

3.2 | Effects of Training Difficulty on Learning Outcome

Results of the Linear Mixed-Effects Model Analysis are provided in Table 4, with correlations between fixed effects in Table 5. The main effect of mean relative difficulty (Estimate: −0.05, 95% CI: [−0.06, −0.04], $p < 0.001$) indicates that, when final student ability is held constant at its mean (0.07), a 1 standard deviation increase in mean relative difficulty is associated with a 5% point decrease in final exam scores. Furthermore, the final exam score was negatively associated with the squared mean relative difficulty (Estimate: −0.03, 95% CI: [−0.04, −0.02], $p < 0.001$), indicating that the relationship between mean relative difficulty and exam scores was curvilinear with a maximum (see Figure 2). This supports the existence of an optimal difficulty level at which exam scores reach a maximum.

3.3 | Differential Effects of Training Difficulty on Learning Outcomes Across Student Abilities

Both final student ability (Estimate: 0.05, 95% CI: [0.04, 0.05], $p < 0.001$) and number of questions taken by medical specialty (Estimate: 2.5×10^{-5}, 95% CI: [2.0×10^{-5}, 3.0×10^{-5}], $p < 0.001$) positively and significantly affect final exam scores, with their effects reflecting relationships with exam performance when the mean relative difficulty is held constant at its average value (−0.09). This shows that controlling for mean relative difficulty during training, students with higher ability and those who train harder tend to achieve superior scores, as expected.

Figure 2 illustrates the model fit, showing the effects of mean relative difficulty (x-axis) and final student ability (colour-coded) on the final exam score (y-axis) across different specialties.

Moving beyond the main effect of final student ability, we found a statistically significant interaction between final student ability and mean relative difficulty (Estimate: −0.02, 95% CI:[−0.03, −0.01], $p < 0.001$). This suggests that the optimal difficulty level significantly varied depending on students' ability as illustrated in Figure 2 by the shifting pattern of vertical lines, colour-coded to represent different ability levels within the same specialty. Although this pattern is not uniform across all specialties, higher-ability students (dark blue) generally have a relatively lower optimal relative difficulty level. Furthermore, there also was an interaction between the squared mean relative difficulty and final student ability (Estimate: −0.01, 95% CI:[−0.01, −0.00], $p < 0.001$), indicating that high-ability students exhibited a greater curvature of the inverted U-shaped relationship between mean question difficulty and final exam scores (as shown by the colour gradient of the curves in Figure 2), and therefore showed greater sensitivity to the mean relative difficulty of questions they were exposed to.

TABLE 4 | Linear mixed-effects model analysis results: Fixed effects.

Fixed effects	Estimates	Std. error	95% CI	*p*
Intercept	0.46	0.02	[0.42, 0.50]	4.72×10^{-11}***
Mean relative difficulty	−0.05	0.01	[−0.06, −0.04]	7.50×10^{-8}***
Mean relative difficulty^2	−0.03	0.01	[−0.04, −0.02]	4.30×10^{-6}***
Final student ability	0.05	0.00	[0.04, 0.05]	2.51×10^{-70}***
Number of questions taken by specialty	2.53×10^{-5}	2.64×10^{-6}	$[2.01\times10^{-5}, 3.05\times10^{-5}]$	1.09×10^{-21}***
Mean relative difficulty: Final student ability	−0.02	0.01	[−0.03, −0.01]	2.68×10^{-4}***
Mean relative difficulty^2: Final student ability	−0.01	0.00	[−0.01, −0.00]	6.43×10^{-5}***
Number of questions taken by specialty: Final student ability	-7.00×10^{-6}	4.84×10^{-6}	$[-1.65\times10^{-5}, 2.48\times10^{-6}]$	1.48×10^{-1}
N specialties	13			
N students	6451			
N observations	45,436			

Note: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

TABLE 5 | Linear mixed-effects model analysis results: Correlation of fixed effects.

Correlation of fixed effects	1	2	3	4	5	6	7	8
1. Intercept	1.00							
2. Mean relative difficulty^2	−0.347	1.00						
3. Mean relative difficulty	−0.406	0.145	1.00					
4. Number of questions taken by specialty	0.011	0.013	0.064	1.00				
5. Final student ability	0.002	0.033	0.361	−0.031	1.00			
6. Mean relative difficulty^2: Final student ability	0.006	−0.003	0.105	0.040	−0.157	1.00		
7. Mean relative difficulty: Final student ability	0.000	0.835	0.052	0.000	0.037	0.137	1.00	
8. Number of questions taken by specialty: Final student ability	−0.004	−0.020	−0.040	−0.487	0.059	−0.050	0.077	1.00

3.4 | Optimal Training Difficulty Differs Between Medical Specialties

Figure 2 demonstrates that the apex of the curves, representing the optimal relative difficulty level for achieving the highest final exam score, varies across specialties. To test the significance of the specialty effect on the apex of the curve, we examined the significance of the random slopes for mean relative difficulty within the specialty grouping factor in our mixed-effects model (Table 6). This was achieved by conducting a likelihood ratio test by using the maximum likelihood (ML) estimation method to compare the full model with a reduced model in which the random slope of mean relative difficulty within the specialty grouping factor was excluded. The comparison, shown in Table 7, reveals a significant improvement in fit upon the inclusion of the random slope of mean relative difficulty within the specialty grouping factor in the Full Model ($\Delta\chi^2 = 339.64$, $\Delta df = 3$, $p < 0.0001$). Additionally, both AIC and BIC are lower for the Full Model, confirming the superior fit. This finding implies that the influence of mean relative difficulty on final exam scores differs between specialties.

As observed in Figure 2, there is a clear trend towards negative mean relative difficulties giving the maximum final exam scores. This pattern is consistent across different medical specialties and varies depending on the final abilities of the students. When adjusted for average final student ability and the number of training questions, the optimal mean relative difficulties across specialties ranged from −1.55 to −0.53, with a mean of −0.952 (SD = 0.378). This suggests that, in this question bank, the best learning outcomes occur when students encounter, on average, slightly easier questions relative to their own ability.

For easier interpretation of our findings, Figure 3 illustrates the correspondence between relative difficulty (x-axis) and

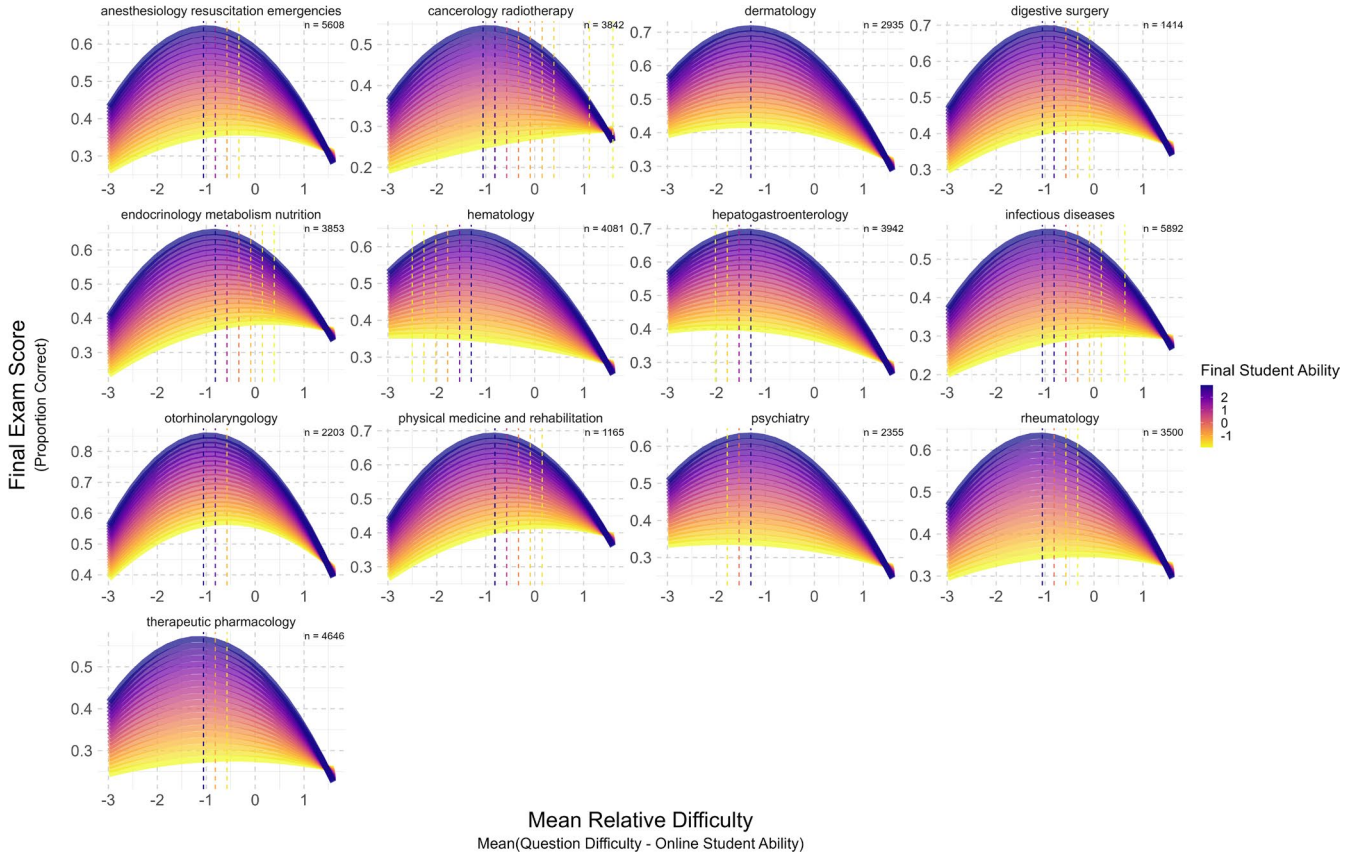


FIGURE 2 | Model fit by specialty: Relation between Mean Relative Difficulty and Final Exam Score (proportion correct) across levels of Final Student Ability, with the Number of questions taken by specialty fixed at the mean value of 280. The colour gradient represents the spectrum of final student abilities, ranging from low (yellow) to high (blue). Vertical dashed lines within each panel, which use the same colour gradient as the curves, indicate the optimal mean relative difficulty (e.g., maxima of the curves) for different levels of final student ability. Additionally, the sample sizes ('n') refer to the number of students included in the model fit for each specialty.

TABLE 6 | Linear mixed-effects model analysis results: Random effects.

Random effects		Variance	Std. Dev.	Corr	
Student	**(Intercept)**	**0.00**	**0.06**		
Specialty	(Intercept)	0.01	0.08		
	Mean relative difficulty	0.00	0.02	−0.47	
	Mean relative difficulty^2	0.00	0.01	−0.88	0.30
Residual		0.01	0.07		

probability of responding correctly (y-axis), using a logistic function transformation of relative difficulty, for various question types (single-answer vs. multiple-answer questions) and numbers of response propositions (ranging from 2–7). The curves show that success probability decreases with relative difficulty following a logistic function. The comparison of the different curves reflects the obvious fact that single-answer questions are easier than multiple-answer questions (which have more possible responses, hence a lower chance level), and that questions with more propositions are more difficult than those with fewer propositions. For multiple-choice questions (MCQs) with five propositions—the most common question type in our question bank—the optimal relative difficulty range, adjusted for average final student ability and the average number of training questions across specialties, falls between −0.53 and −1.55 (indicated by red vertical lines). This range corresponds to a success probability of 0.64 to 0.83 (shaded area).

4 | Discussion

In this study focused on medical learning, we investigated the impact of training difficulty on learning outcomes.

To address our main research question (RQ1), we exploited the random allocation of training questions to medical students to use question difficulty as an instrumental variable. For each student and specialty, we computed the mean difference between question difficulty and online student ability during training (mean relative difficulty) and examined its influence on learning outcomes. Controlling for students' final ability in the medical specialty and the number of questions taken by specialty, our analysis reveals a significant quadratic

TABLE 7 | Comparison of linear mixed models.

Model	npar	AIC	BIC	logLik	Deviance	$\Delta\chi^2$	Δdf	p
Reduced Model	13	−96,307	−96,194	48,167	−96,333	—	—	—
Full Model	16	−96,641	−96,501	48,336	−96,673	339.64	3	< 0.0001***

Note: p-values indicate the significance of the improvement in fit. Significance Codes: '***' $p < 0.001$, '**' $p < 0.01$, '*' $p < 0.05$.

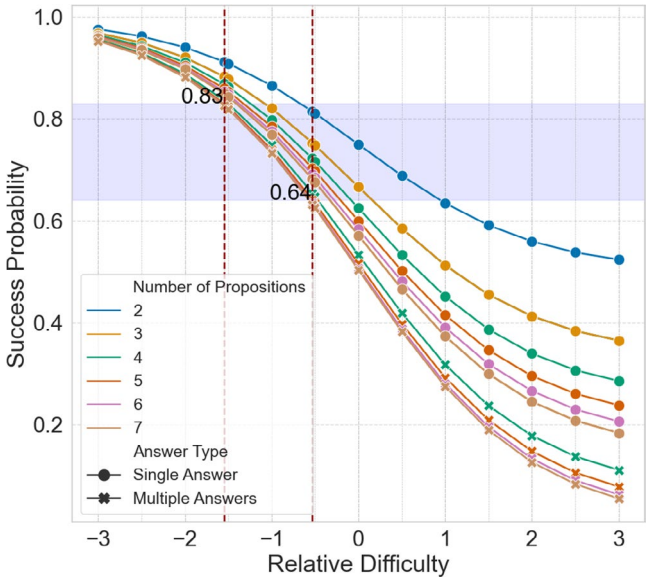


FIGURE 3 | Correspondence between relative difficulty and success probability for questions of different types and number of propositions. The shaded area represents the optimal success probability range corresponding to a relative difficulty range of −0.53 and −1.55 for the most frequent 5-proposition multiple-answer questions with numerical values marking the boundaries.

relationship between mean relative difficulty and exam score. This finding supports the inverted *U*-shaped hypothesis, indicating the existence of an optimal difficulty level in this particular medical training setting. In our investigation of the second research question (RQ2), we analysed the optimal difficulty level and its variation across different medical specialties and final student abilities. We found that the optimal difficulty level for achieving the best learning outcomes depends on both the final student's ability and the medical specialty in question.

In order to better appreciate the size of this inverted U-shape effect, one might consider a student of average ability (0.07) taking questions at his/her level, thereby experiencing a mean relative question difficulty of 0, in one particular specialty, say, anesthesiology (upper left in Figure 2). This corresponds to being correct about 50% of the time in multiple-answer questions (as shown in Figure 3). Following this training, this student would have an expected score of about 45% at the final anesthesiology exam (from Figure 2). By changing his training regime to relatively easier questions (a relative difficulty of −1, corresponding to answering correctly 70% of the time), his expected score would improve to 47% (following the orange line on Figure 2). However, due to the interaction between student and linear and quadratic mean difficulty, a student of higher ability (maximum student ability = 2.93, the dark blue line in Figure 2) would experience a greater benefit of shifting his training questions from a relative difficulty of 0 to −1: his predicted score would rise from 59% to 64%. Whereas a student at a lower ability level (minimum student ability = −1.92, yellow line) would actually see his predicted score decrease slightly by training on even easier questions. Finally, this pattern differs depending on specialty, as one can see in Figure 2 and as confirmed by the significance of the specialty random slope. These effects of a substantial change in training question difficulty on predicted exam scores may seem relatively small. However, it should be recalled that the specific characteristics of our dataset could only lead to minimising effect sizes. Indeed, students trained both inside and outside the platform, but we had no information on the latter. Our analysis was restricted to the influence of training that was observable on the platform, potentially missing most of the training for some students. Furthermore, we didn't have any experimental control over training question difficulty. Thus, students experienced very variable question difficulties, and the effect of mean question difficulty was limited by randomly occurring variation between students. Finally, outcomes were measured across broad medical specialties, assessing the far transfer of knowledge from training to the ECN exam, rather than rote memorization. Other educational settings where all the training occurs on the platform, where question difficulty is entirely controlled, and where the knowledge domains are narrower should expect larger effect sizes.

Our primary finding of an optimal difficulty level aligns with the intuitive understanding of many fundamental pedagogical principles, suggesting that effective learning occurs when tasks strike a balance between being neither too easy nor too difficult. Such an intuition has been formulated in prominent educational theories such as "zone of proximal development" by Vygotsky (1978), and Principles of Instruction by Rosenshine (2012). Although quantification of success depends on task structure and chance level, our optimal success probability range of 64%–83% is compatible with the 85% rule for optimal learning in stochastic gradient-descent-based learning algorithms identified in the simulation work of Wilson et al. (2019). This optimal success probability of 64%–83% (for 5-proposition multiple-choice questions) that we observed may suggest that students learned best from very easy questions. This may actually reveal that the questions in the BNE medical training system are excessively difficult on average, and that adding easier questions, or adaptively adjusting question difficulty to final student ability would significantly benefit students. Implementing such changes would provide a more personalised and effective learning experience, consistent with prior research that highlights the benefits of tailoring training content (Zou et al. 2019; Maghsudi et al. 2021; Benvenuti et al. 2023).

Our results also conflict with some previous research. For instance, Jansen et al. (2013), in their study on the impact of training difficulty on math performance in children aged 8 to

13 years, found that the easiest difficulty level led to the most effective learning. They used a computer-adaptive program to adjust the difficulty of math problems to individual abilities, targeting success rates of 60%, 75%, and 90% for difficult, medium, and easy conditions, respectively. Furthermore, they identified a mediating effect of the number of problems attempted. Specifically, children in easier conditions attempted more problems in their platform, which indirectly led to greater math improvement. This mediation effect could explain the linear relationship between success rate and final performance observed in their study while our quasi-experimental study revealed a more curvilinear relationship. Additionally, the nature of the tasks in their study differed from ours. Their tasks required a single correct answer, which meant that success rates were straightforward to measure and control. In contrast, our study involved questions with multiple possible answers, adding complexity to the success rate measurement. This difference in task structure may contribute to the contrasting results between their study and ours.

Additionally, Lomas et al. (2013) reported in their controlled experiment involving a basic educational game that simpler problems increased engagement but resulted in a slower rate of learning. They had initially hypothesized that moderate levels of challenge would maximize engagement, following the inverted-*U* hypothesis. They attributed these unexpected findings to their game not aligning with the typical close game hypothesis proposed by Abuhamdeh and Csikszentmihalyi (2012), where the challenge is motivating when there is high uncertainty about winning or losing. Importantly, their study differed from ours as it involved a different learning task and lacked explicit feedback after each question, thus missing the integration of uncertainty about winning or losing. Thus, there may be no such thing as a universal optimal difficulty level or consistent applicability of the inverted-*U* shape hypothesis across all learning scenarios. Rather, optimal training conditions may depend on the learning domain, on the type of training, as well as on students' abilities and characteristics.

Indeed, we found evidence for such variability in our own results, with optimal difficulty levels ranging from −1.55 (about 83% success rate) in haematology to −0.53 (about 64%) in physical medicine and rehabilitation, with a mean of −0.952 (SD = 0.378) (see Table 8). This may reflect the state of the current BNE question bank, with questions in different specialties being designed by different professors, leading to specialties with questions of different difficulty levels on average.

Furthermore, even within a given specialty, optimal difficulty levels depended on the final student's ability. Higher-ability students tended to have more pronounced negative curvature, thus showing greater sensitivity to the mean relative difficulty, while lower-ability students tended to have flatter response curves. The interpretation is not straightforward, but this may in part be due to some lower-ability students responding at random regardless of question difficulty, thus being less sensitive to this factor.

Furthermore, in all but 4 specialties (out of 13), the optimal difficulty level for the best students was lower than that for the lower-ability ones. This may seem paradoxical, but one should recall that this is a *relative* question difficulty (absolute question difficulty—online student ability). Thus, to take digestive surgery in Figure 2 as an example, an optimal difficulty level of −1 for students with an ability of 2 still represents a much greater absolute question difficulty (1) than an optimal difficulty level of 0 for students with an ability of −2 (−2). So it is the case that better students benefit from more difficult questions (in an absolute sense), but the optimal difference between question difficulty and student ability seems to decrease with ability. Why the opposite trend is observed in 3 specialties (haematology, hepatogastroenterology, and psychiatry) is not clear, and would first warrant replication before any attempt at an interpretation.

TABLE 8 | Optimal relative difficulty by medical specialty for the most frequent 5-proposition multiple-answer questions.

Medical specialty	Optimal relative difficulty
Anesthesiology, Resuscitation, Emergencies	−0.8099
Cancerology, Radiotherapy	−0.5785
Dermatology	−1.3190
Digestive Surgery	−0.7174
Endocrinology, Metabolism, Nutrition	−0.5323
Haematology	−1.5504
Hepatogastroenterology	−1.5042
Infectious Diseases	−0.6248
Otorhinolaryngology	−0.8099
Physical Medicine and Rehabilitation	−0.5323
Psychiatry	−1.4116
Rheumatology	−0.9488
Therapeutic Pharmacology	−1.0413

In terms of implications, our findings emphasise the importance of customising practice exercises to individual needs. While many systems achieve this adaptation by dynamically adjusting difficulty levels based on past performance or by externally determining an optimal level, they often lack a firm empirical grounding. In the BNE system, it should now be possible to enhance learning outcomes by selecting questions at precisely the optimal level given a student's current ability in a given medical specialty.

Our study differs from existing research in several notable respects. First, unlike prior studies that often focus solely on simple declarative knowledge, our platform encompasses a comprehensive array of medical knowledge spanning a six-year medical university curriculum. This complexity is reflected in our final exam, which assesses learning outcomes with an entirely new set of questions designed to evaluate students' ability to apply medical knowledge to unfamiliar scenarios. This differs from many other studies that rely on repeated questions within a much narrower knowledge domain.

Second, while some studies utilise educational games, our platform is specifically designed for training purposes, with the primary goal being compulsory learning rather than entertainment. This distinction may be important, as factors influencing intrinsically motivated activities may differ from those affecting non-intrinsically motivated activities (Elliot and Harackiewicz 1994; Zhang and Yu 2022). Additionally, our analysis involves an extensive training period lasting 4 months, a duration notably longer than many other studies in the field.

Finally, unlike similar studies in the field that rely on learning curves established by reference questions (Papoušek et al. 2016b), we used learning outcomes that were measured through the scores of students in a final exam. This offered an ecologically valid external validation of learning.

Overall, all these factors contribute to the robustness and uniqueness of the present study.

4.1 | Limitations

There are several notable limitations in our current study. First, it is important to acknowledge that our findings are specifically tied to medical learning within the context of the chosen educational platform. The observed inverted *U*-shaped relationship for optimal difficulty, as demonstrated in our study, adds to the existing literature documenting this phenomenon across various disciplines, including digital gaming (Abuhamdeh and Csikszentmihalyi 2012), online chess (Abuhamdeh et al. 2015), and literacy training (Ronimus et al. 2014). Furthermore, our findings contribute to the well-established evidence supporting the positive impact of personalising the difficulty level of learning content based on students' behaviours and performance (Major et al. 2021), reinforcing its relevance in adaptive learning systems. This cross-disciplinary consistency suggests that our findings may have broader applicability. Of course, none of the specific results and predictions reported here can be straightforwardly translated to other learning contexts, since they are specific to medical training in this cohort; indeed they even vary by specialty, and they also depend on the specific set of questions available on the platform. Rather, our results suggest that similar modelling or experimental manipulation of relative question difficulty in other learning contexts may produce similar effects, and thus it may be possible to determine an optimal difficulty level for each student in each learning context. Still, further research is required to explore the generalizability of these results across diverse educational contexts and subject areas.

Second, while previous studies often employ controlled conditions (Papoušek et al. 2016b) or allow users to freely select the difficulty levels (Lomas et al. 2013), this study could not experimentally manipulate relative question difficulty in a randomised controlled design, which would have been the most powerful approach to address our research questions. However, we were able to exploit the random drawing of questions to adopt a quasi-experimental design in observational data. This is an equally valid way to estimate causal effects when direct intervention is not possible, which is frequently used in epidemiology and in economics.

More generally, contrary to experimental settings, we had no control whatsoever over the students' learning experience. Students were entirely free to use this platform for training or not: some did so very intensively, others very little. They were also free to choose specialties and types of questions (isolated questions, progressive files) to train on. Most students likely had other learning opportunities outside the platform, and so to different degrees. We had no information whatsoever on training performed outside the platform; we only had access to a subset of their overall training experience. The one parameter that was not controlled by students was question difficulty, whose effect we were, therefore, able to estimate. However, our estimation of relative question difficulty only concerns the part of their training that students did on the platform and ignores the difficulty of training done outside the platform. Thus, our evaluation of the effect of relative question difficulty must be underestimated.

The last limitation is the lack of engagement analysis in our study, due to constraints within our dataset. Existing literature suggests that difficulty levels may impact learning and engagement differently (Papoušek et al. 2016b). Thus, incorporating engagement analysis could provide valuable insights into the relationship between difficulty, learning outcome, and student engagement.

4.2 | Perspectives

The current study and its identified limitations prompt new research questions. Firstly, it would be desirable to re-assess our research questions using much more controlled settings, with learners' training experience entirely documented, and ideally using a randomised controlled design to experimentally manipulate relative question difficulty.

Building upon the identification of optimal difficulty levels for various medical specialties and student abilities, it would seem logical to implement an adaptive learning algorithm within the educational platform, which would dynamically present questions at the appropriate difficulty level for each learner in each specialty. It would then be important to assess whether overall learning outcomes improve compared to the non-adaptive version of the platform.

While our study offers insights into the relationship between difficulty and learning outcomes, it leaves unanswered questions regarding student motivation and engagement. Recent research has yielded mixed findings in this area: while one study found no clear effect of difficulty–skill balance on engagement (Cutting et al. 2023), another (Israel-Fishelson and Hershkovitz 2021) reported a positive association between task difficulty and micro-persistence—defined as the tendency to successfully complete individual tasks, a nuanced measure of engagement. These mixed results underscore the need for further investigation into how difficulty influences engagement, aiming to achieve a balanced approach that offers exercises not too hard, and not too boring.

Furthermore, a recent study (Lomas et al. 2017) discovered that when users were allowed to choose their own difficulty

levels, moderately challenging tasks were the most motivating. Conversely, when difficulty was predetermined, the easiest tasks were found to be the most motivating. This suggests exploring the possibility of leaving learners the choice of training difficulty level. It would then be interesting to evaluate whether this has the intended effect on motivation, and furthermore whether this also has a positive effect on learning.

5 | Conclusion

This study provides significant insights into the effects of training difficulty on learning outcomes, supporting the Inverted *U*-shape Hypothesis, which posits that there is an optimal level of difficulty that maximizes learning. Furthermore, it demonstrates that this optimal level of difficulty varies according to the subject matter and the learner's proficiency in that subject. These insights have important implications for the education and learning analytics communities, advocating for the adjustment of training difficulty in an adaptive and personalised way to ensure that all learners encounter optimally challenging content, tailored to their individual learning curves and the specific demands of the subjects they are studying.

6 | Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author(s) used Grammarly and ChatGPT in order to improve the readability and language of the manuscript. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Author Contributions

Erva Nihan Kandemir: conceptualization, data curation, formal analysis, software, visualization, writing – original draft, methodology. **Jill-Jênn Vie:** writing – review and editing, validation, software, formal analysis. **Adam Sanchez-Ayte:** writing – review and editing, software. **Olivier Palombi:** writing – review and editing, validation. **Franck Ramus:** conceptualization, formal analysis, supervision, validation, writing – review and editing, project administration, methodology.

Acknowledgements

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

The data necessary to reproduce the database creation and the analyses is not publicly accessible for reasons of privacy for the participants.

The analytic code necessary to reproduce the analyses presented in this paper is publicly accessible: Python and R scripts are available at https://github.com/ervanihank/optimal_difficulty_medicine_students.git.

Funding

This work was supported by Agence Nationale de la Recherche, ANR-10-IDEX-0001-02, ANR-17-EURE-0017, ANR-21-CE28-0030.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data necessary to reproduce the database creation and the analyses are not publicly accessible for reasons of privacy for the participants. The analytic code (Python and R scripts) necessary to reproduce the analyses presented in the paper is publicly accessible. To enhance transparency, the markdown output generated from these analyses is also available, enabling readers to trace the analysis process without direct access to the data.

References

Abdi, S., H. Khosravi, and S. Sadiq. 2021. Modelling Learners in Adaptive Educational Systems: A Multivariate Glicko-Based Approach. In LAK21: 11th International Learning Analytics and Knowledge Conference, 497–503.

Abuhamdeh, S., and M. Csikszentmihalyi. 2012. "The Importance of Challenge for the Enjoyment of Intrinsically Motivated, Goal-Directed Activities." *Personality and Social Psychology Bulletin* 38: 317–330.

Abuhamdeh, S., M. Csikszentmihalyi, and B. Jalal. 2015. "Enjoying the Possibility of Defeat: Outcome Uncertainty, Suspense, and Intrinsic Motivation." *Motivation and Emotion* 39: 1–10.

Antal, M. 2013. "On the Use of Elo Rating for Adaptive Assessment." *Studia Universitatis Babeș-Bolyai, Informatica* 58: 29–41.

Attali, Y. 2014. "A Ranking Method for Evaluating Constructed Responses." *Educational and Psychological Measurement* 74: 795–808.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67: 1–48.

Benvenuti, M., A. Cangelosi, A. Weinberger, et al. 2023. "Artificial Intelligence and Human Behavioral Development: A Perspective on New Skills and Competences Acquisition for the Educational Context." *Computers in Human Behavior* 148: 107903.

Cao, Y., S.-Y. Gong, Z. Wang, Y. Cheng, and Y.-Q. Wang. 2022. "More Challenging or More Achievable? The Impacts of Difficulty and Dominant Goal Orientation in Leaderboards Within Educational Gamification." *Journal of Computer Assisted Learning* 38: 845–860.

Chan, W.-L., and D.-Y. Yeung. 2021. Clickstream Knowledge Tracing: Modeling How Students Answer Interactive Online Questions. In LAK21: 11th International Learning Analytics and Knowledge Conference, 99–109.

Cutting, J., S. Deterding, S. Demediuk, and N. Sephton. 2023. "Difficulty-Skill Balance Does Not Affect Engagement and Enjoyment: A Pre-Registered Study Using Artificial Intelligence-Controlled Difficulty." *Royal Society Open Science* 10: 220274.

De Morais, A. M., J. M. Araujo, and E. B. Costa. 2014. "Monitoring Student Performance Using Data Clustering and Predictive Modelling." In *In 2014 IEEE Frontiers in Education Conference (FIE) Proceedings*, 1–8. IEEE.

Deterding, S., and J. Cutting. 2023. "Objective Difficulty-Skill Balance Impacts Perceived Balance but Not Behaviour: A Test of Flow and Self-Determination Theory Predictions." *Proceedings of the ACM on Human-Computer Interaction* 7: 1179–1205.

Elliot, A. J., and J. Harackiewicz. 1994. "Goal Setting, Achievement Orientation, and Intrinsic Motivation: A Mediational Analysis." *Journal of Personality and Social Psychology* 66: 968–980.

Elo, A. E. 1978. *The Rating of Chess Players Past and Present*. Arco Publishing.

García-Pérez, M. A. 1998. "Forced-Choice Staircases With Fixed Step Sizes: Asymptotic and Small-Sample Properties." *Vision Research* 38: 1861–1881.

Gersten, R. M., D. W. Carnine, and P. B. Williams. 1982. "Measuring Implementation of a Structured Educational Model in an Urban School

District: An Observational Approach." *Educational Evaluation and Policy Analysis* 4: 67–79.

Good, T. L., and D. A. Grouws. 1977. "Teaching Effects: A Process-Product Study in Fourth-Grade Mathematics Classrooms." *Journal of Teacher Education* 28: 49–54.

Gweon, G.-H., H.-S. Lee, C. Dorsey, R. Tinker, W. Finzer, and D. 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.

Hattie, J. 2012. *Visible Learning for Teachers: Maximizing Impact on Learning*. Routledge.

Israel-Fishelson, R., and A. Hershkovitz. 2021. "Micro-Persistence and Difficulty in a Game-Based Learning Environment for Computational Thinking Acquisition." *Journal of Computer Assisted Learning* 37: 839–850.

Jansen, B. R., J. Louwerse, M. Straatemeier, S. H. Van der Ven, S. Klinkenberg, and H. L. Van der Maas. 2013. "The Influence of Experiencing Success in Math on Math Anxiety, Perceived Math Competence, and Math Performance." *Learning and Individual Differences* 24: 190–197.

Kandemir, E. N., J.-J. Vie, A. Sanchez-Ayte, O. Palombi, and F. Ramus. 2024. Adaptation of the Multi-Concept Multivariate Elo Rating System to Medical Students' Training Data. In Proceedings of the Fourteenth International Conference on Learning Analytics and Knowledge (LAK 2024).

Kim, J. H. 2019. "Multicollinearity and Misleading Statistical Results." *Korean Journal of Anesthesiology* 72: 558–569.

Klinkenberg, S., M. Straatemeier, and H. L. J. 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: 1813–1824.

Kostons, D., T. van Gog, and F. Paas. 2010. "Self-Assessment and Task Selection in Learner-Controlled Instruction: Differences Between Effective and Ineffective Learners." *Computers & Education* 54: 932–940.

Kuznetsova, A., P. B. Brockhoff, and R. H. B. Christensen. 2020. lmerTest: Tests in Linear Mixed Effects Models. https://CRAN.R-project.org/package=lmerTest.

Lin, L.-C., I.-C. Hung, Kinshuk, and N.-S. Chen. 2019. "The Impact of Student Engagement on Learning Outcomes in a Cyber-Flipped Course." *Educational Technology Research and Development* 67: 1573–1591.

Lomas, D., K. Patel, J. L. Forlizzi, and K. R. Koedinger. 2013. "Optimizing Challenge in an Educational Game Using Large-Scale Design Experiments." In *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 89–98. ACM.

Lomas, J. D., K. Koedinger, N. Patel, S. Shodhan, N. Poonwala, and J. L. Forlizzi. 2017. Is Difficulty Overrated? the Effects of Choice, Novelty and Suspense on Intrinsic Motivation in Educational Games. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 1028–1039.

Lüdecke, D., D. Makowski, M. S. Ben-Shachar, P. Waggoner, and I. Patil. 2021. Performance: Assessment of Regression Models Performance R Package Version 0.12.4. https://easystats.github.io/performance/.

Lüdecke, D. 2021. sjPlot: Data Visualization for Statistics in Social Science. R package Version 2.8.15. https://CRAN.R-project.org/package=sjPlot.

Ma, W., O. O. Adesope, J. C. Nesbit, and Q. Liu. 2014. "Intelligent Tutoring Systems and Learning Outcomes: A Meta-Analysis." *Journal of Educational Psychology* 106: 901–918.

Maghsudi, S., A. Lan, J. Xu, and M. van Der Schaar. 2021. "Personalized Education in the Artificial Intelligence Era: What to Expect Next." *IEEE Signal Processing Magazine* 38: 37–50.

Major, L., G. A. Francis, and M. Tsapali. 2021. "The Effectiveness of Technology-Supported Personalised Learning in Low-And Middle-Income Countries: A Meta-Analysis." *British Journal of Educational Technology* 52: 1935–1964.

Papoušek, J., and R. 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.

Papoušek, J., R. Pelánek, and V. Stanislav. 2014. Adaptive Practice of Facts in Domains With Varied Prior Knowledge. In Educational Data Mining 2014.

Papoušek, J., V. Stanislav, and R. Pelánek. 2016a. Evaluation of an Adaptive Practice System for Learning Geography Facts. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, 134–142.

Papoušek, J., V. Stanislav, and R. Pelánek. 2016b. "Impact of Question Difficulty on Engagement and Learning." In *Intelligent Tutoring Systems: 13th International Conference, ITS 2016, Zagreb, Croatia, June 7–10, 2016. Proceedings*, vol. 13, 267–272. Springer International Publishing.

Park, O. C., and J. Lee. 2003. "Adaptive Instructional Systems." *Educational Technology Research and Development* 25: 651–684.

Pelánek, R. 2016. "Applications of the Elo Rating System in Adaptive Educational Systems." *Computers & Education* 98: 169–179.

Pelánek, R., J. Papoušek, J. Řihák, V. Stanislav, and J. Nižnan. 2017. "Elo-Based Learner Modeling for the Adaptive Practice of Facts." *User Modeling and User-Adapted Interaction* 27: 89–118.

Piech, C., J. Bassen, J. Huang, et al. 2015. "Deep Knowledge Tracing." In *Advances in Neural Information Processing Systems (NIPS)*, 505–513. Curran Associates, Inc.

Power, J. 2019. "The Influence of Task Difficulty on Engagement, Performance and Self-Efficacy." In *Explorations in Technology Education Research: Helping Teachers Develop Research Informed Practice*, 157–169. Springer.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Education Research.

Roediger, H. L., III, and J. D. Karpicke. 2006. "The Power of Testing Memory: Basic Research and Implications for Educational Practice." *Perspectives on Psychological Science* 1: 181–210.

Romero, C., S. Ventura, E. L. Gibaja, C. Hervás, and F. Romero. 2006. "Web-Based Adaptive Training Simulator System for Cardiac Life Support." *Artificial Intelligence in Medicine* 38: 67–78.

Ronimus, M., J. Kujala, A. Tolvanen, and H. Lyytinen. 2014. "Children's Engagement During Digital Game-Based Learning of Reading: The Effects of Time, Rewards, and Challenge." *Computers & Education* 71: 237–246.

Rosenshine, B. 2012. "Principles of Instruction: Research-Based Strategies That All Teachers Should Know." *American Educator* 36: 12.

Salden, R. J., F. Paas, and J. J. Van Merriënboer. 2006. "Personalised Adaptive Task Selection in Air Traffic Control: Effects on." *Human Factors* 48: 120–130.

Sampayo-Vargas, S., C. J. Cope, Z. He, and G. J. Byrne. 2013. "The Effectiveness of Adaptive Difficulty Adjustments on Students' Motivation and Learning in an Educational Computer Game." *Computers & Education* 69: 452–462.

Schütt, A., T. Huber, J. Nasir, C. Conati, and E. André. 2024. "Does Difficulty Even Matter? Investigating Difficulty Adjustment and Practice Behavior in an Open-Ended Learning Task." *Proceedings of the 14th Learning Analytics and Knowledge Conference*: 253–262.

Sheather, S. 2009. *A Modern Approach to Regression With R*. Springer Science & Business Media.

Sinha, T., P. Jermann, N. Li, and P. Dillenbourg. 2014. "Your Click Decides Your Fate: Inferring Information Processing and Attrition

Behavior From Mooc Video Clickstream Interactions." *arXiv Preprint arXiv*: 1407.7131.

van der Linden, W. J., and R. K. Hambleton. 2013. *Handbook of Modern Item Response Theory*. Springer Science & Business Media.

Vygotsky, L. 1978. "Interaction Between Learning and Development." *Readings on the Development of Children* 23: 34–41.

Wauters, K., P. Desmet, and W. Van den Noortgate. 2010. "Adaptive Item-Based Learning Environments Based on the Item Response Theory: Possibilities and Challenges." *Journal of Computer Assisted Learning* 26: 549–562.

Wauters, K., P. Desmet, and W. Van Den Noortgate. 2011. Monitoring Learners' Proficiency: Weight Adaptation in the Elo Rating System. In EDM, 247–252.

Westlin, J., E. A. Day, and M. G. Hughes. 2019. "Learner-Controlled Practice Difficulty and Task Exploration in an Active-Learning Gaming Environment." *Simulation & Gaming* 50: 812–831.

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. https://ggplot2.tidyverse.org.

Wickham, H., R. François, L. Henry, and K. Müller. 2023. dplyr: A Grammar of Data Manipulation. https://CRAN.R-project.org/package=dplyr.

Wilson, R. C., A. Shenhav, M. Straccia, and J. D. Cohen. 2019. "The Eighty Five Percent Rule for Optimal Learning." *Nature Communications* 10: 4646.

Zhang, Q., and Z. Yu. 2022. "Meta-Analysis on Investigating and Comparing the Effects on Learning Achievement and Motivation for Gamification and Game-Based Learning." *Education Research International* 2022: 1519880.

Zhang, Y., and W.-B. Goh. 2021. "Personalized Task Difficulty Adaptation Based on Reinforcement Learning." *User Modeling and User-Adapted Interaction* 31: 753–784.

Zou, X., W. Ma, Z. Ma, and R. S. Baker. 2019. "Towards Helping Teachers Select Optimal Content for Students." In *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25–29, 2019, Proceedings, Part II 20*, 413–417. Springer.