Article

Time Saving Students' Formative Assessment: Algorithm to Balance Number of Tasks and Result Reliability

Jaroslav Melesko * and Simona Ramanauskaite

Department of Information Technologies, Vilnius Gediminas Technical University, Sauletekio al. 11, LT-10223 Vilnius, Lithuania; simona.ramanauskaite@vilniustech.lt

* Correspondence: jaroslav.melesko@vilniustech.lt

Featured Application: The proposed algorithm may be applied in computerized adaptive testing for formative assessment, particularly in a bring your own device setting with a cloud hosted formative assessment tool. The algorithm can reduce the number of items that have to be solved by a student and provide a faster way to determine a suitable personalized learning path for students. The algorithm identifies the weakest areas of students' knowledge and indicates the easiest items from which to start corrective learning. Given the universal spread of mobile devices and access to cloud resources by educators, we believe that the proposed formative assessment method may be a valid alternative to traditional assessment approaches and IRT systems.

Abstract: Feedback is a crucial component of effective, personalized learning, and is usually provided through formative assessment. Introducing formative assessment into a classroom can be challenging because of test creation complexity and the need to provide time for assessment. The newly proposed formative assessment algorithm uses multivariate Elo rating and multi-armed bandit approaches to solve these challenges. In the case study involving 106 students of the Cloud Computing course, the algorithm shows double learning path recommendation precision compared to classical test theory based assessment methods. The algorithm usage approaches item response theory benchmark precision with greatly reduced quiz length without the need for item difficulty calibration.

Keywords: formative assessment; multi-armed bandit; multivariate Elo-rating; upper-confidence bound algorithm; personalized learning; e-learning; adaptive testing

Citation: Melesko, J.; Ramanauskaite, S. Time Saving Students' Formative Assessment: Algorithm to Balance Number of Tasks and Result Reliability. *Appl. Sci.* **2021**, *11*, 6048. <https://doi.org/10.3390/app11136048>

Academic Editor: Miguel Ángel Conde

Received: 5 June 2021

Accepted: 28 June 2021

Published: 29 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Formative assessment is a crucial component of the learning process, because it measures student's knowledge, maps the path towards learning goals, and tracks progress [1,2]. This process quickens and makes learning more effective [3]. In addition, recent studies indicate that the use of formative assessment is associated with more feelings of autonomy and competence, and more autonomous motivation in students [4]. In contrast, summative assessment often carries punitive connotations, because students are penalized for their mistakes with a lower grade [5]. Instead of passing judgment on student's knowledge level, the purpose of formative assessment is to provide constructive feedback in order to accelerate learning [6,7], frame learning goals, and monitor progress towards them [8]. According to Gareis [9], this can be accomplished by taking into account three factors that make formative assessment effective. Firstly, formative assessment is most effective when it is used to tailor student's learning paths based on their current level of understanding. Secondly, results of formative assessment should not affect student's evaluation, as they do in the case of summative assessment. This helps students take ownership of their learning process and focus on acquiring new knowledge. Thirdly, formative assessment should be used to provide feedback and guide students towards their learning goals [10]. There is recent evidence that pedagogical benefits of formative assessment are even more

pronounced in online, cloud based, and asynchronous environments [11]. According to Marzano et al. feedback is the essential pedagogic technique for improving student's performance [12]. Studies indicate that feedback may improve learning outcomes by up to 0.4 standard deviation [13]. However, these results are still being debated [14]. A recent study observed an increase in academic achievement without a noticeable increase in motivation in students, which is often attributed to feedback from formative assessments [15]. One of the factors that contributes to the quality of feedback is how well it addresses the fundamental pedagogic need of closing the gap between current and desired student proficiencies [16]. Another factor is the delay between the measurement of student's knowledge and feedback, which can be days or even weeks in case of manually graded tests. This delay may negatively affect the positive effects of feedback [17]. Finally, feedback in a form of a mark does not provide a path towards increased proficiency [18], and thus does not address one of the main functions of formative assessment and feedback [19].

Time spent on formative assessment must be taken into account when considering its practical value. Students' knowledge is constantly changing, as consequence students must be repeatedly reassessed [20]. This is time consuming. At the same time, if the student will be forced to demonstrate the knowledge he or she already mastered, the motivation to do the formative assessment might be lost. Therefore, it is necessary to assure the balance between the needed number of tasks and the assessment accuracy (weakest areas of student's knowledge and the best learning path).

The paper proposes an algorithm capable of reducing formative assessment length while conforming to good formative assessment practices. This computer adaptive testing method may be sufficiently simple to use in a classroom setting with minimal time expenditure. The algorithm does not require pretesting, this could enable educators to host the assessment tool on a cloud and have students partake in formative assessment. While mobile devices may be less convenient than full sized paper and computer tests, greatly reduced by the proposed algorithm quiz length may mitigate this inconvenience.

2. Need of New Formative Assessment Algorithms

Classical test theory (CTT) has a long history of use in test construction and utilization [21], it is based on the assumption that tests measure a sum of a true score and error score. While latent trait theory (LTT), also known as item response theory (IRT) takes a more nuanced approach such as statistical models which take into account different item difficulty and the existence of multidimensional latent trait space [22,23]. It is based on the idea that the probability of a correct response to an item is a function of the personal latent traits of a student and an item [24,25]. This approach is often used in computerized adaptive testing [26,27]. In adaptive testing items presented to a student are chosen based on previously seen items and student's responses. This approach can shorten test times and precision by, for example, maximizing information gain from each item [28,29]. However, IRT based tests require item pool calibration to determine item parameters used in the calculation of correct answer probability [30,31]. This increases test creation and iterative improvement complexity. In practical applications, this complexity may inhibit educators in deploying IRT based formative assessment tests.

Elo rating based systems can perform similar roles to IRT based solutions. Elo rating was originally developed for matching chess players of similar skills [32]. The same principles can be used to estimate item difficulty, proficiency of students, and matching the two in the educational setting [33–35]. Both, IRT and Elo rating systems usually utilize the logistic function to determine the probability of correct answer to an item. However, Elo rating based systems determine item parameters dynamically, each answered item's difficulty setting is immediately adjusted during the test. This removes the need for a separate calibration step before the test deployment. As with IRT, multivariate, measuring more than one latent trait, Elo rating systems have been proposed [36,37].

There are other approaches to model students and knowledge acquisition [38], these include performance factor analysis [39,40], knowledge tracing [41–43], and multi-arm

bandit (MAB) solutions [44–46]. Multi-armed bandit algorithms are named after a gambler's problem, where they have to choose which slot machine, one armed bandit, to play in order to maximize their reward based on previous pay-off observations [47–49]. These algorithms offer a way of negotiating exploration and exploitation dilemma, searching for best paying slot machines and exploiting best paying slot machines. This problem arises in many settings, such as education, advertisement placement, website optimization, and packet routing [50–52].

In this paper, a hybrid testing algorithm based on Elo rating and upper confidence bound (UCB) multi-armed bandit approach is proposed and applied in the educational setting. It combines the advantages of Elo rating systems, such as dynamic calibration of item difficulties, guesses parameters and compatibility with latent trait theory, with an efficient exploration-exploitation trade-off solution. The algorithm is applied in the context of formative assessment, to explore student's knowledge and suggest further learning material in accordance with Hattie's feedback model [16], which is still a dominant model for formative assessment [53]. Firstly, the instructor forms an item pool divided into topics or competencies, "Where am I going?" Secondly, students partake in formative assessment where questions are chosen and dynamically calibrated by the proposed algorithm, "How am doing?" Thirdly, the algorithm can reveal with weakest areas of knowledge and suggest the easiest items, learning objects, within that area of knowledge to learn, "Where to next?" The proposed algorithm can significantly reduce test creation complexity compared to IRT solutions and reduce test length. Unlike CCT based tests the algorithm takes in account varying item complexity, item guesses probabilities, and produces twice more accurate item suggestions than classical methods and previously proposed UCB algorithm [45].

3. Materials and Methods

3.1. Research Methodology

As stated in the introduction, the main function of formative assessment is to provide students with feedback [12]. The quality of feedback hinges on locating a student in knowledge space and providing a route towards the desired competencies [16]. To solve this challenge and take into account the difficulty of the tasks, the proposed algorithm simultaneously measures item difficulties and student's multivariate proficiencies as it explores item pools. Several components are needed to apply and evaluate the algorithm: a way to model and generate synthetic students, a way to assess data from real students, components that will perform assessment (testing), benchmark estimation component and components for measuring assessment performance against a baseline. The schematic representation of the utilized system is presented in Figure 1.

3.2. Formative Assessment Algorithms

3.2.1. Elo-Bandit Algorithm

The focus of the section is the newly proposed in this paper algorithm based on Elo rating and upper confidence bound algorithms, further referenced to as the Elo-bandit algorithm (EBA). It builds on a less precise, classical test theory [45]. The problem of formative assessment can be modeled as taking an item pool on several topics as input, performing the assessment of students, and returning individualized learning suggestions for each student. Suggestions should reflect the topic in most dire need of instruction and item difficulty (easier items first). During the assessment, the algorithm has to negotiate exploration of topics on which student knowledge is yet unknown and investigate weak topics ("exploitation" in multi-armed bandit terminology) to precisely determine which items are known to a student. To evaluate student's knowledge on each topic and assess item difficulty the proposed EBA algorithm uses Elo rating system.

Figure 1. Schematic representation of research methodology and its principal components.

In simple terms, UCB algorithm first chooses the topic of the next question. Two components go into this choice, already demonstrated knowledge on the topic (Elo estimation) and uncertainty of that knowledge (for example, if a student answered only one item from the topic, the estimation of their knowledge is unlikely to be accurate). The weaker the demonstrated knowledge and the higher the uncertainty the more likely the topic will be chosen. The emergent behavior is that the algorithm at first tries to explore all the topics and then gradually focuses on areas in greatest need of instruction. Because those topics ought to be corrected post assessment and gaining detailed knowledge about them is pedagogically valuable. Simultaneously, as all the students answer the questions the algorithm updates their difficulty estimation. Every correct answer lowers question difficulty, every incorrect answer increases it. However, the magnitude of the change depends on the knowledge of the question recipient: if the knowledgeable on the topic student answers an item the negative change will be small. Because the student is expected to know the answers even to hard questions and it does not necessarily mean question difficulty estimation was inaccurate. The magnitude of the change also decreases as more data is gathered. Algorithm 1 shows the proposed EBA algorithm.

Algorithm 1. The proposed Elo-bandit algorithm for test item selection.

1. // Find unused items i for each topic T with non-zero item-topic weights
 2. **For each topic do**
 3. **For each item do**
 4. **If** $w_{it} \neq 0$ **and** $i \notin U_s$ **then**
 5. $t \cup \{i\}$
 6. // Calculate lower confidence bound of estimated Elo b_t
 7. **For each** $t \neq \emptyset$ **do**
 8. $b_t = sElo_t - C \sqrt{\frac{\log n}{N_t}}$
 9. Choose a topic with the smallest b_t value $T_c = \underset{t \in T}{\operatorname{argmin}}(b_t)$
 10. Choose the item with the smallest Elo from T_c , $i_c = \underset{i \in T_c}{\operatorname{argmin}}(Elo_i)$
 11. Add the item to the list of seen items by the student, $U_s \cup \{i_c\}$
 12. Calculate the probability of correct answer, $P_c = P_g + \frac{1 - P_g}{1 + 10^{\frac{Elo_i - sElo_t}{400}}}$
 13. // Update item and student knowledge Elo estimates based on the answer
 14. **For each topic do**
 15. $sElo'_t = sElo_t + w_{it}k(X_{si} - P_c)$
 16. $Elo'_i = Elo_i - j(X_{si} - P_c)$
-

Firstly, the algorithm attempts to select the least familiar topic, then asks the easiest question within that topic. The weakest topic selection is performed using the following equation:

$$t_n = \underset{t \in T}{\operatorname{argmin}} \left(sElo_t - C \sqrt{\frac{\log n}{N_t}} \right) \quad (1)$$

where t is a chosen topic for a quiz question n , with $sElo_t$ difficulty estimation. The second parameter in brackets describes uncertainty in difficulty estimation and is controlled by exploration constant C , the total number of questions asked n , and the number of questions asked on the topic N_t . The uncertainty parameter decreases as the algorithm probes student's knowledge on the topic and $sElo_t$ estimation become more accurate [54]. Each student is assigned a starting Elo rating on each topic t , $sElo_t = 1000$.

The algorithm takes as input a pool of items, with weights describing their belonging to a number of topics and the guess probability factor. For example, a multiple-choice quiz item with one correct answer option out of four has a guess factor of $P_g = 0.25$. If no item difficulty calibration data is provided, EBA then assigns each item an initial Elo rating $iElo = 1000$. The probability of correct answer P_c to an item with $iElo$ score on a single topic t can be calculated using a shifted logistic function:

$$P_c = P_g + \frac{1 - P_g}{1 + 10^{\frac{iElo - sElo_t}{400}}} \quad (2)$$

The base of the exponent, and the denominator in the exponent are the same as in Elo score used in chess matchmaking. Where they were chosen for human readability. The new student's Elo rating on a given topic $sElo'_t$ after a student gives either correct, $X_{ans} = 1$, or incorrect, $X_{ans} = 0$, answer to an item is calculated based on the expected probability of correct answer:

$$sElo'_t = sElo_t + k(X_{ans} - P_c) \quad (3)$$

Constant k in Equation (3) specifies the sensitivity of Elo rating to change, or the maximum change Elo rating can make given the most unlikely outcome. If k is small, the Elo estimation converges slowly, if k is large then Elo score oscillates with large amplitude around the true score without ever reaching it. A common solution is to use a decaying constant, which converges fast on true value when there are few data points to estimate Elo rating, but ultimately settles on a more accurate value than a large k value. There are more mathematically rigorous methods of modeling uncertainty inherent in the Elo

rating estimation and using it as a basis for the constant, but they do not produce superior results [55]. Convergence of item Elo rating estimation $iElo$ and student's topic Elo rating $sElo$ towards a value estimated using Maximum Likelihood Estimation (MLE) method can be seen in Figure 2.

Figure 2. Item Elo rating estimation approaching MLE estimation, on the left. On the right, student's topic knowledge estimation.

Seed Elo used to generate synthetic data is also plotted in the figure. A more detailed description of seed Elo and synthetic student generation is presented in the following sections.

To summarize, the algorithm probes each student's topic knowledge and simultaneously estimates item difficulties. Then this knowledge can be used to tailor feedback for each student. In this paper, we chose to select the easiest items from the weakest topic for a student, as this would allow a student to progress from easier to hard content and possibly motivate them. However, the differently inclined instructor may choose another approach, for example selecting items with Elo score best matching student's topic knowledge. Thus, post assessment each student can receive not just a mark and a set of items which they got wrong, but a number of areas where their knowledge is weakest and a way to correct these deficiencies starting from the simplest items.

3.2.2. Other Formative Assessment Algorithms

The Elo-bandit algorithm proposed in this article is compared to the previously presented upper confidence bound based algorithm, and two classical formative assessment methods, choosing questions at random and sequential enumeration through topics.

The upper confidence bound algorithm identifies and explores the weakest topic within the quiz material, without taking in account item difficulty or the order in which those items ought to be studied to improve topic knowledge [45]. UCB algorithm uses the same principle to select a topic as EBA, Equation (1). Except, it is based on classical test theory and uses a mean answer value instead of Elo rating estimation to denote student's knowledge of the topic. In practice, UCB based algorithm tends to first explore all topics, as exploration parameter starts large. As the number of questions answered on a particular topic increases, so the uncertainty parameter decreases and the mark on the topic decides which topic shall be further explored by the algorithm.

The random algorithm simply selects a random question out of the pool of yet unseen by the student items. The enumerate topics algorithm serves a random question from each unexhausted topic in sequence. Its operation resembles UCB when its operation is dominated by the exploration component.

3.3. Data Preparation for Algorithm Validation

3.3.1. Simulated Students' Data

Before applying the proposed in this paper algorithm to assess real students, a number of simulations were performed. The number of synthetic students in each experiment was set to 120, to be close to an expected number of real students. A large number would give EBA an advantage, as it would mean that each question item would receive more answers from students simplifying its difficulty estimation procedure. Each simulated assessment had items on multiple topics. Each topic was represented by a vector of weights equal in length to the number of items within the quiz. Each element of the topic vector denoted the item's relevancy to the topic. The number of items on each topic was set to be equal when possible. When the number of items was indivisible by the number of topics, the remaining items were assigned to separate topics.

The difficulty of simulated items, seed Elo, was generated using a normal distribution with mean, $\mu = 1000$, and standard deviation, $\sigma = 300$. Each synthetic student was assigned a general level of knowledge $sElo$ from a normal distribution, $\mu = 1000$ and $\sigma = 400$. Then that value was used as a basis for topic knowledge $sElo_t$, $\mu = sElo$ and $\sigma = 200$. Finally, the student's capacity to answer an item was calculated by estimating relevant to the item Elo rating using, item weights, w_{ti} , and topic knowledge:

$$aElo_{si} = \sum_{t=1}^T sElo_t * w_{ti} \quad (4)$$

Then the correct answer was chosen as a Bernoulli random variable with the probability calculated using Equation (2), guess factor $P_g = 0.25$.

3.3.2. Real Students' Data

Two datasets were used to test the presented algorithms. A newly presented 2021 dataset was from a formative assessment of real students which was carried using Moodle system. The data was collected using a quiz covering 15 topics of a cloud computing course. The course topics varied from common knowledge on computer networks to cloud types, platforms, and cloud services. The size of the item pool was 80 items. Students had 80 min to finish the quiz, and it took them 39 min on average to do it. The quiz was held on 11 February 2021 and had no summative impact on the final grade. In total 106 Vilnius Gediminas Technical University, undergraduate students participated and agreed to share their data for scientific analysis. Each student was served all items from the pool and their answers were recorded in a comma separated file. The file was then used as a proxy for students to test different algorithms. This way each algorithm received the same input data. All 80 items were multiple choice questions with one correct answer, guess factor $P_g = 0.25$. Algorithm accuracy was measured in relation to item difficulty and student's topic knowledge as estimated by MLE based IRT with the entire dataset as input. While algorithms had to operate on incomplete data, first asking each student one question, then two, and so on.

The second dataset explored in this article was already used in publication [45]. This dataset was included to also measure Elo-Bandit algorithm against previously presented upper confidence bound based algorithm using same data and same classical test theory based benchmarks. Data was collected in a quiz held at Vilnius Gediminas Technical University, Lithuania, on 25 April 2019. It had 104 undergraduate student participants, who each answered 60 items covering 15 topics. In the case of CTT based experiment topic knowledge of students was assumed to be the mean of all answers on said topic.

3.4. Validation Measurements and Metrics

Item response theory was used with parameters calculated using iterative maximum likelihood estimation method to provide a benchmark against which the performance of algorithms is measured. Benchmark values represent idealized IRT results if a full length

IRT test was applied to calibration data. MLE method maximizes likelihood function so that observed data is most probable. It begins with some a priori value for a student's knowledge on each topic and item difficulty Elo ratings. Then student's knowledge ratings are recalculated to best reflect observed quiz outcome. Afterward item difficulty ratings are recalculated in light of newly estimated student's topic knowledge. This process was iteratively repeated until the mean change in Elo ratings fell below one point.

For synthetic students and quizzes, for whom seed Elo rating values used for a generation were known, the mean error between MLE and seed Elo rating was calculated and is presented in Figure 3. The number of topics within the quiz was set to 15. The quiz length in the graph on the right was set to 80 items. The number of students in the graph on the right is 120. These numbers were chosen to resemble the parameters of the experiment with real student data. However, it is important to note that, synthetic student generation parameters may be significantly different from latent trait distribution in real students.

Figure 3. Mean error of MLE calculated values as compared to seed values used in the generation of synthetic students. On the left mean error is a function of group size. On the right mean error as a function of quiz length.

Assessment of accuracy for LTT based experiments was measured against IRT MLE values. An argument could be made that the item difficulty should be personalized for each student, rather than relying on statistical data as in the case of IRT. However, this would require a more sophisticated psychometric theory than IRT and other approaches such as learning styles are heavily criticized [56,57]. As for IRT, latent values were not used because they are unknown for real students and without reference to real values synthetic scores can be manipulated: if the deviation in topic knowledge is small, separating best known topics from worst will be harder. The set of suitable items according to IRT for a student was established by selecting topics least familiar to a student (lowest $sElo_t$ values) using complete knowledge of quiz answers, then easiest items (lowest $iElo$ values) from those topics were added to the set. Delta values for the selection of suitable items were chosen so that an average of 15 items were chosen for a student in synthetic and real experiments. This number was chosen for two reasons. Items in real quizzes mostly covered a narrow band of material. Selecting a few of them would mean students would quickly amend the gaps in their knowledge and would need to either retake formative assessment or study without guidance. Thus, wasting time. In this setting, a hypothetical average student with equal knowledge on every topic would have 1 item in need of studying from each topic of the course. Additionally, the number of suitable items depends on the proficiency of a student. If the number is chosen too low formative assessment will fail to provide a sufficient amount of material to the most knowledgeable students.

Once the set of suitable items for each student was known assessment performance was measured using precision, or positive predictive value (PPV). PPV describes the probability of providing suitable learning material to a student based on data gathered by formative assessment. PPV is calculated, $PPV = TP / (TP + FP)$. Where TP is true positive, the number of items in the set of suitable items according to a formative assessment

algorithm at some length of a quiz. TP is an intersection of the set of MLE selected suitable items and algorithm selected items. TP + FP (false positive) is simply the number of elements in the set of items selected by the algorithm.

Accuracy assessment for CTT based experiments targeting weakest topic prediction was also measured using precision metric. Except, item difficulty was not taken into account, and the student's topic knowledge was measured by a mark for that topic. The median number of suitable topics for a student is set to one, to be consistent with the previous study.

4. Results

4.1. LTT Item Precision Simulation

Formative assessment precision in establishing personalized suitable items for each synthetic student is presented in Figure 4 as measured against IRT MLE values. It presents mean precision $\pm$ sample standard deviation for 10 groups of students taking 10 different quizzes. Measurements were performed for 10 synthetic groups with 120 students each.

Figure 4. Formative assessment precision selecting items as a function of quiz length for synthetic students. Simulated data from 10 groups, taking 10 different quizzes, 120 students in each group.

The Elo-bandit algorithm performs significantly better than UCB and classical test methods: random and enumerate. These methods quickly approach 50% precision as compared to IRT, but never achieve 60% or higher precision. While EBA achieves 80% precision by question 21, and 90% by 45, as marked in the plot. No algorithm, however, achieves ideal IRT accuracy, but Elo-bandit comes close to ideal MLE based IRT assessment with 98% precision. Simulated results depend on the assumptions about values and distribution of latent traits, which only approximate real students.

4.2. CTT Topic Precision Empirical Data

The same algorithms were compared using real student data from the 2019 dataset (104 students) to measure their precision in selecting topics using classical test theory Figure 5. The figure presents the measured precision of suitable topic detection dependence on quiz length involving real students' results are presented as mean precision $\pm$ standard error of the mean (SEM).

The Elo-bandit has virtually the same performance as the UCB based algorithm in this application. Even though it breaches the 95% precision mark sooner, it lags in reaching 70% accuracy. Both of these algorithms offer about a 50% reduction in quiz length

compared to classical testing methods when searching for the weakest topics. Enumerate and random algorithms perform similarly, enumerate having a better performance in medium quiz lengths.

Figure 5. Formative assessment precision selecting topics in CCT as a function of quiz length, 2019 dataset. A vertical dashed line denotes 95% precision.

A more detailed look at the precision distribution for UCB and Elo-bandit algorithms for the 2019 dataset CCT model is presented in Figure 6.

Figure 6. Distribution of students assessed for topic precision for UCB, left, and Elo-bandit, right. Where red <1%, orange 1–50%, dark green 50–99%, and green >99% precision.

Both algorithms display very similar behavior. The Elo-bandit shows slightly fewer students with wildly incorrect, less than one percent precision, estimates at quiz lengths shorter than 30 questions. This is because UCB uses the average score to measure topic knowledge. This estimate overshoots real value with a low sample size, while Elo rating estimation approaches the final estimate more smoothly. However, UCB demonstrates fewer students in the 50–99% category past 30 question mark, because it uses the same mean score as the benchmark ground truth CCT measure. Therefore, given full information, it will always reach full agreement with the benchmark.

4.3. LTT Item Precision Empirical Data

The latent trait theory model was applied in estimating item suggestions on real student data, 2021 dataset Figure 7.

Figure 7. Mean algorithm precision estimating personalized suitable items as measured against IRT MLE.

A formative assessment quiz included 80 items on 15 topics and was taken by 106 students. Results confirm observations made using synthetic students. The Elo-bandit is a promising alternative to IRT based tests, achieving almost the same precision 97% with half the test length. The Elo-bandit has significantly better precision finding suitable items than CCT based algorithms. The Elo-bandit algorithm shows better performance on real data compared to simulation. Offering more than twice the precise results for tests longer than 15 questions. The 80% precision was achieved at 15 questions in comparison to 23 questions for simulation, and 90% precision at 23 in contrast to 45 question simulation value. Despite the fact that constants for the algorithm were calibrated using simulated data. A possible explanation is provided in the following discussion section.

Question difficulties and Pearson coefficient correlation matrix for the 2021 dataset are presented in Figure 8.

Figure 8. Question difficulties in 2021 dataset on the left and item correlation matrix on the right.

The question difficulties graph shows a percentage of correct answers each item has received. A number of items were answered correctly by less than 25% of students, despite having a guess factor of $P_g = 0.25$. Colored bars illustrate selected suitable items for a sample strong student: dark green—true positive, red—false positive, yellow—false negative, light green—true negative. The correlation matrix demonstrates low generally correlation even for inter topic questions except for items 64–69. Strongly correlated items

would improve the Elo-bandit and UCB algorithms' performance. Probing an item from a strongly correlated topic brings more information about other topic items, so fewer questions are needed to identify familiar and unfamiliar subjects for a student.

5. Discussion

The Elo-bandit algorithm shows promise as an alternative for IRT based and CTT based tests in formative assessment applications. Despite the pedagogical value formative assessments provide [2,18], they have an opportunity cost: assessments consume time that could be spent directly studying course material through lectures and self-study. Time spent on each activity has a demonstrable effect on learning outcomes [58]. In addition to time savings, shorter assessment length may combat decision fatigue, possibly leading to more efficient post-test learning [59]. The Elo-bandit algorithm provides an opportunity to benefit from formative assessment while saving time and mitigating fatigue. Potential quiz length reduction observed in the experimental setting was about 50%. In addition, informative assessment precision is less important than summative assessment. Imprecision in summative assessment may lead to failure to obtain employment or admission to the university to a deserving candidate. While a degree of imprecision in formative assessment may cause a teaching system to serve a suboptimal learning object. Therefore, it might be rational to aim at lower precisions such as 80 or 90 percent when it allows for significant quiz length reduction. In the presented 2021 case study with an item pool consisting of 80 questions median time of assessment amongst 106 students was 39 min or about 30 s per item. Settling for 80% or 90% precision would allow to cut median assessment time to less than 8 and 12 min respectively. Such reduction in quiz length may allow instructors to incorporate formative assessment in the learning process where previously it was impractical. From an educational point of view pedagogical value of a method is a function of more than just its mathematical properties. The viability of the method also depends on pedagogical scaffolding around it and, of course, student psychology. To gauge the algorithm's impact on learning outcomes further experiments would be needed that are beyond the scope of this paper. Perhaps the key factor limiting IRT based assessment in classrooms is the need for calibration [26]. A teacher cannot simply write up and add items to a question bank as they teach. In addition to requiring a large student pool, unavailable to many teachers, item difficulty parameters are also dependent on the student background exacerbating the problem. The proposed in this paper algorithm does not require calibration, to the best of my knowledge no similar formative assessment solution have been proposed and applied using data from real students. However, Elo rating has been used to estimate learning object difficulty. In 2019 Mangaroska et al. published data from 22 students indicating positive results in the programming exercise recommendation system [35]. Pankiewicz and Bator found Elo rating to be a superior task difficulty measure to proportion correct and learner feedback [60]. Finally, based on simulations Antal has found that Elo rating may be preferable to IRT for adaptive item sequencing in practice systems as it does not require item pretesting [61]. These findings are in agreement with positive results obtained in this study: Elo rating is a viable alternative to IRT assessment in a formative context. Of course, item difficulties are subjective to each student, as everyone possesses unique competencies and talents. However, this effect can be lessened if a group of students being assessed has a similar academic background. This is easier to achieve for methods that do not require pre-testing, as the group being tested itself becomes the reference of difficulty.

Simulations demonstrated significant algorithm performance differences in comparison with real data implying that generation parameters used for synthetic students did not reflect students in the classroom. The algorithm's accuracy gets worse with the decrease in item difficulty and correlation. Real items were largely uncorrelated, but more difficult, resulting in shorter quiz lengths. This also means that Elo-bandit assessment is most effective when used for unfamiliar material as is often the case for formative assessment, a mean number of correct answers for real students was $\mu = 40$. However, the proposed algorithm may be impractical if applied in a context of summative assessment without

modifications or if used for material that is very familiar to students. Another setting where the Elo-bandit algorithm may be less effective is in small groups and with new items. Because the algorithm does not require pre-testing and uses answers during the quiz to estimate item difficulties, the smaller the group the less accurate estimate may be acquired. This shortcoming may be mitigated by maintaining item difficulty parameters across assessments or manually guessing starting item difficulties when they are added to an item bank. Finally, topic size has an effect on the advantage in length reduction conveyed by the exploration-exploitation mechanic, as was explored in the paper on UCB formative assessment [45]. To summarize, measuring topic level precision, with item difficulty ignored, the Elo-bandit algorithm did not demonstrate increased performance to justify increased complexity over UCB algorithm but was still twice as quick as classical testing methods. In the context where item difficulty and precise item level suggestions are needed CCT based algorithms cannot replace IRT and Elo based assessment methods without further modifications, as their measurement accuracy was twice worse even at full quiz length. The Elo-bandit algorithm may also offer formative quiz length reduction in learning environments where item pre-testing is not feasible. However, evaluating the pedagogical value of the proposed algorithm will require experiments that measure its impact on the learning process in different settings.

6. Conclusions and Future Work

The proposed Elo-bandit algorithm can reduce formative assessment time by up to 70% when targeting 90% item level precision. This feature is student friendly as it saves them time in comparison to traditional methods, as well as saving them from fatigue. The proposed algorithm is also beneficial to instructors because it does not require item pre-testing. These advantages may enable instructors to introduce cloud hosted formative assessment tools accessible through mobile devices in a class setting.

The synthetic data, used in preliminary experiments, did not accurately match data from real students. As the assumed distribution of item difficulties and student proficiencies did not match with the one observed in empirical data. However, the proposed Elo-bandit algorithm does not require normal distribution of item difficulties and demonstrated better performance with real data. This shows the algorithm's flexibility and potential for practical value in real classrooms. The current version of the Elo-bandit algorithm is designed to be brief and identify the weakest topics and easiest items for short-term learning path generation. In the future, it could be extended to generate a long-term learning path and retain a model of student's knowledge between assessments. This would reduce assessment length even further and mitigate the need for repeated assessments providing even greater benefits to students in the form of saved time and mental resources. However, to measure the algorithm's value as a pedagogical tool in real and virtual classrooms more experiments on learning outcomes, student satisfaction and its viability for teaching different subjects with varying item difficulty distributions and topics would be needed. These and other factors will ultimately impact the proposed algorithm's adoption as a formative assessment tool.

Author Contributions: Conceptualization, J.M.; methodology, J.M.; software, J.M.; validation, J.M.; formal analysis, J.M. and S.R.; investigation, J.M.; data curation, J.M.; writing—original draft preparation, J.M.; writing—review and editing, S.R.; visualization, J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Anonymized data was gathered in accordance with legislation as a byproduct of the normal study process. Experiments performed on datasets had no impact on subjects.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data available on request.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Black, P.; Wiliam, D. Assessment and Classroom Learning. *Assess. Educ. Princ. Policy Pract.* **1998**, *5*, 7–74. [\[CrossRef\]](#)
- Dunn, K.E.; Mulvenon, S.W. A Critical Review of Research on Formative Assessments: The Limited Scientific Evidence of the Impact of Formative Assessments in Education. *Pract. Assess. Res. Eval.* **2009**, *14*, 7.
- Guskey, T. Formative Classroom Assessment and Benjamin S. Bloom: Theory, Research, and Implications. In Proceedings of the Annual Meeting of the American Educational Research Association, Montreal, QC, Canada, 11–15 April 2005.
- Leenknecht, M.; Wijnia, L.; Köhlen, M.; Fryer, L.; Rikers, R.; Loyens, S. Formative assessment as practice: The role of students' motivation. *Assess. Eval. High. Educ.* **2021**, *46*, 236–255. [\[CrossRef\]](#)
- Scriven, M. The methodology of evaluation. In *Social Science Education Consortium*; Purdue University: West Lafayette, IN, USA, 1967.
- Sadler, D.R. Formative Assessment: Revisiting the territory. *Assess. Educ. Princ. Policy Pract.* **2007**, *5*, 77–84. [\[CrossRef\]](#)
- Moss, C.M.; Brookhart, S.M. *Advancing Formative Assessment in Every Classroom: A Guide for Instructional Leaders*; ASCD: Alexandria, VA, USA, 2019; ISBN 1416626727.
- Tomasik, M.J.; Berger, S.; Moser, U. On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Front. Psychol.* **2018**, *9*, 2245. [\[CrossRef\]](#)
- Group, A.R. *Assessment for Learning: Beyond the Black Box*; Qualifications and Curriculum Authority: Coventry, UK, 1999; ISBN 0856030422.
- Sadley, D.R. Formative assessment and the design of instructional systems. *Instr. Sci.* **1989**, *18*, 119–144. [\[CrossRef\]](#)
- McLaughlin, T.; Yan, Z. Diverse delivery methods and strong psychological benefits: A review of online formative assessment. *J. Comput. Assist. Learn.* **2017**, *33*, 562–574. [\[CrossRef\]](#)
- Marzano, R.; Pickering, D.; Pollock, J. *Classroom Instruction That Works: Research-Based Strategies for Increasing Student Achievement*; Pearson Education Ltd.: Harlow, Essex, UK, 2004; ISBN 987-0131195035.
- Shute, V.J. Focus on formative feedback. *Rev. Educ. Res.* **2008**, *78*, 153–189. [\[CrossRef\]](#)
- Bennett, R.E. Formative assessment: A critical review. *Assess. Educ. Princ. Policy Pract.* **2011**, *18*, 5–25. [\[CrossRef\]](#)
- Ozan, C.; Kincal, R.Y. The effects of formative assessment on academic achievement, attitudes toward the lesson, and self-regulation skills. *Educ. Sci. Theory Pract.* **2018**, *18*, 85–118. [\[CrossRef\]](#)
- Hattie, J.; Timperley, H. The Power of Feedback. *Rev. Educ. Res.* **2007**, *77*, 81–112. [\[CrossRef\]](#)
- Gareis, C.R. Reclaiming an important teacher competency: The lost art of formative assessment. *J. Pers. Eval. Educ.* **2007**, *20*, 17–20. [\[CrossRef\]](#)
- Guskey, T.R.; Bailey, J.M. *Developing Grading and Reporting Systems for Student Learning*; Corwin Press: Thousand Oaks, CA, USA, 2001; ISBN 080396854X.
- Nicol, D.; Macfarlane-Dick, D. *Rethinking Formative Assessment in HE: A Theoretical Model and Seven Principles of Good Feedback Practice*; The Higher Education Academy: York, UK, 2006.
- Choi, Y.; McClenen, C. Development of adaptive formative assessment system using computerized adaptive testing and dynamic bayesian networks. *Appl. Sci.* **2020**, *10*, 8196. [\[CrossRef\]](#)
- Novick, M.R. The axioms and principal results of classical test theory. *J. Math. Psychol.* **1966**, *3*, 1–18. [\[CrossRef\]](#)
- Hambleton, R.K.; Swaminathan, H.; Cook, L.L.; Eignor, D.R.; Gifford, J.A. Developments in Latent Trait Theory: Models, Technical Issues, and Applications. *Rev. Educ. Res.* **1978**, *48*, 467–510. [\[CrossRef\]](#)
- Meiser, T.; Plieninger, H.; Henninger, M. IRTree models with ordinal and multidimensional decision nodes for response styles and trait-based rating responses. *Br. J. Math. Stat. Psychol.* **2019**, *72*, 501–516. [\[CrossRef\]](#) [\[PubMed\]](#)
- Reckase, M.D. The past and future of multidimensional item response theory. *Appl. Psychol. Meas.* **1997**, *21*, 25–36. [\[CrossRef\]](#)
- Kolen, M.J. Comparison of traditional and item response theory methods for equating tests. *J. Educ. Meas.* **1981**, *18*, 1–11. [\[CrossRef\]](#)
- Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*; Routledge: London, UK, 1980; ISBN 1136557245.
- Jia, J.; Le, H. The design and implementation of a computerized adaptive testing system for school mathematics based on item response theory. In *International Conference on Technology in Education*; Springer: Singapore, 2020.
- Wang, W.; Song, L.; Wang, T.; Gao, P.; Xiong, J. A Note on the Relationship of the Shannon Entropy Procedure and the Jensen–Shannon Divergence in Cognitive Diagnostic Computerized Adaptive Testing. *SAGE Open* **2020**, *10*. [\[CrossRef\]](#)
- McDonald, A.S. The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Comput. Educ.* **2002**, *39*, 299–312. [\[CrossRef\]](#)
- Syeda, Z.F.; Shahzadi, U.; Ali, G. Rasch Calibration of Achievement Test: An Application of Item Response Theory. *SJESR* **2020**, *3*, 426–432. [\[CrossRef\]](#)
- Edelen, M.O.; Reeve, B.B. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual. Life Res.* **2007**, *16*, 5–18. [\[CrossRef\]](#) [\[PubMed\]](#)
- Elo, A.E. *The Rating Of Chess Players, Past & Present*; Ishi Press: San Jose, CA, USA, 2008; ISBN 978-0923891275.
- Pelánek, R. Applications of the Elo rating system in adaptive educational systems. *Comput. Educ.* **2016**, *98*, 169–179. [\[CrossRef\]](#)

34. Abdi, S.; Khosravi, H.; Sadiq, S. Modelling learners in adaptive educational systems: A multivariate glicko-based approach. In Proceedings of the Lak21: 11th International Learning Analytics and Knowledge Conference, Virtual. Irvine, CA, USA, 12–16 April 2021. [\[CrossRef\]](#)
35. Mangaroska, K.; Vesin, B.; Giannakos, M. Elo-rating method: Towards adaptive assessment in e-learning. In Proceedings of the IEEE 19th International Conference on Advanced Learning Technologies, ICALT 2019, Maceio, Brazil, 15–18 July 2019.
36. Abdi, S.; Khosravi, H.; Sadiq, S.; Gasevic, D. A multivariate elo-based learner model for adaptive educational systems. *arXiv* **2019**, arXiv:1910.12581.
37. Doebler, P.; Alavash, M.; Giessing, C. Adaptive experiments with a multivariate Elo-type algorithm. *Behav. Res. Methods* **2015**, *47*, 384–394. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Desmarais, M.C.; Baker, R.S.J.D. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Model. User Adapt. Interact.* **2012**, *22*, 9–38. [\[CrossRef\]](#)
39. Pavlik, P.I.; Cen, H.; Koedinger, K.R. Performance factors analysis—A new alternative to knowledge tracing. In Proceedings of the 14th International Conference on Artificial Intelligence in Education, Brighton, UK, 6–10 July 2009.
40. Liu, M.; Kitto, K.; Shum, S.B. Combining factor analysis with writing analytics for the formative assessment of written reflection. *Comput. Hum. Behav.* **2021**, *120*, 106733. [\[CrossRef\]](#)
41. Corbett, A.T.; Anderson, J.R. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Model. User Adapt. Interact.* **1994**, *4*, 253–278. [\[CrossRef\]](#)
42. Wu, J.; Huang, Z.; Liu, Q.; Lian, D.; Wang, H.; Chen, E.; Ma, H.; Wang, S. Federated Deep Knowledge Tracing. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM 21, Virtual event. Israel, 8–12 March 2021.
43. Zhou, Y.; Li, X.; Cao, Y.; Zhao, X.; Ye, Q.; Lv, J. LANA: Towards Personalized Deep Knowledge Tracing Through Distinguishable Interactive Sequences. *arXiv* **2021**, arXiv:2105.06266.
44. Liu, Y.; Mandel, T.; Brunskill, E.; Popovi, Z. Trading Off Scientific Knowledge and User Learning with Multi-Armed Bandits. In Proceedings of the 7th International Conference on Educational Data Mining EDM14, London, UK, 4–7 July 2014.
45. Melesko, J.; Novickij, V. Computer adaptive testing using upper-confidence bound algorithm for formative assessment. *Appl. Sci.* **2019**, *9*, 4303. [\[CrossRef\]](#)
46. Lin, F. Adaptive quiz generation using Thompson sampling. In Proceedings of the Third Workshop Eliciting Adaptive Sequences for Learning (WASL 2020), Cyberspace. 6 July 2020.
47. Lai, T.; Robbins, H. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **1985**, *6*, 4–22. [\[CrossRef\]](#)
48. Agrawal, R. Sample mean based index policies by $O(\log n)$ regret for the multi-armed bandit problem. *Adv. Appl. Probab.* **1995**, *27*, 1054–1078. [\[CrossRef\]](#)
49. Vermorel, J.; Mohri, M. Multi-armed bandit algorithms and empirical evaluation. In Proceedings of the 16th European Conference on Machine Learning, ECML 2015, Porto, Portugal, 3–7 October 2005.
50. Bubeck, S.; Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.* **2012**, *5*, 1–122. [\[CrossRef\]](#)
51. Clement, B.; Roy, D.; Lopes, M.; Oudeyer, P.; Clement, B.; Roy, D.; Lopes, M.; Optimization, P.O.O. Online Optimization and Personalization of Teaching Sequences. In Proceedings of the DI: Digital Intelligence—1st International Conference on Digital Cultures, Nantes, France, 17–19 September 2014.
52. Lan, A.S.; Baraniuk, R.G. A Contextual Bandits Framework for Personalized Learning Action Selection. In Proceedings of the 9th International Conference on Educational Data Mining EDM16, Raleigh, NC, USA, 29 June–2 July 2016; pp. 424–429.
53. Panadero, E.; Andrade, H.; Brookhart, S. Fusing self-regulated learning and formative assessment: A roadmap of where we are, how we got here, and where we are going. *Aust. Educ. Res.* **2018**, *45*, 13–31. [\[CrossRef\]](#)
54. Szepesvari, C.; Lattimore, T. *Bandit Algorithms*; Cambridge University Press: Cambridge, UK, 2020; ISBN 978-1108486828.
55. Nižnan, J.; Pelánek, R.; Rihák, J. Student Models for Prior Knowledge Estimation. In Proceedings of the 8th International Conference on Educational Data Mining EDM15, Madrid, Spain, 26–29 June 2015.
56. Kirschner, P.A. Stop propagating the learning styles myth. *Comput. Educ.* **2017**, *106*, 166–171. [\[CrossRef\]](#)
57. Papadatou-Pastou, M.; Gritzali, M.; Barrable, A. The Learning Styles Educational Neuromyth: Lack of Agreement Between Teachers’ Judgments, Self-Assessment, and Students’ Intelligence. *Front. Educ.* **2018**, *3*. [\[CrossRef\]](#)
58. Bratti, M.; Staffolani, S. Student Time Allocation and Educational Production Functions. *Ann. Econ. Stat.* **2013**, *103*. [\[CrossRef\]](#)
59. Pignatiello, G.A.; Martin, R.J.; Hickman, R.L. Decision fatigue: A conceptual analysis. *J. Health Psychol.* **2020**, *25*, 123–135. [\[CrossRef\]](#)
60. Pankiewicz, M.; Bator, M. Elo Rating Algorithm for the Purpose of Measuring Task Difficulty in Online Learning Environments. *e-Mentor* **2019**, *82*, 43–51. [\[CrossRef\]](#)
61. Antal, M. On the use of ELO rating for adaptive assessment. *Stud. Inform.* **2013**, *58*, 29–41.