applied sciences

MDPI

Article

Computer Adaptive Testing Using Upper-Confidence Bound Algorithm for Formative Assessment

Jaroslav Melesko [1,*] and Vitalij Novickij [2]

[1] Department of Information Technologies, Vilnius Gediminas Technical University, LT-10223 Vilnius, Lithuania

[2] Institute of High Magnetic Fields, Vilnius Gediminas Technical University, LT-10223 Vilnius, Lithuania; vitalij.novickij@vgtu.lt

* Correspondence: jaroslav.melesko@vgtu.lt

Received: 11 September 2019; Accepted: 11 October 2019; Published: 14 October 2019

Featured Application: The paper proposes an application of UCB algorithm for item selection in formative assessment. The main advantage of this approach is its ease of implementation when compared to Elo and Multidimensional Item Response Theory based testing. Thus the method should be applicable in virtually any classroom where formative assessment is desired and students have access to computers or phones.

Abstract: There is strong support for formative assessment inclusion in learning processes, with the main emphasis on corrective feedback for students. However, traditional testing and Computer Adaptive Testing can be problematic to implement in the classroom. Paper based tests are logistically inconvenient and are hard to personalize, and thus must be longer to accurately assess every student in the classroom. Computer Adaptive Testing can mitigate these problems by making use of Multi-Dimensional Item Response Theory at cost of introducing several new problems, most problematic of which are the greater test creation complexity, because of the necessity of question pool calibration, and the debatable premise that different questions measure one common latent trait. In this paper a new approach of modelling formative assessment as a Multi-Armed bandit problem is proposed and solved using Upper-Confidence Bound algorithm. The method in combination with e-learning paradigm has the potential to mitigate such problems as question item calibration and lengthy tests, while providing accurate formative assessment feedback for students. A number of simulation and empirical data experiments (with 104 students) are carried out to explore and measure the potential of this application with positive results.

Keywords: formative assessment; Upper-Confidence Bound algorithm; Multi-Armed Bandit algorithm; e-Learning; intelligent tutoring systems; education

1. Introduction

Formative assessment have been proposed to make education more accessible and more effective [1–4]. The distinction between summative and formative roles of assessment was first proposed by Scriven [5] and then applied to students by Bloom [6,7]. Formative assessment is specifically intended to generate feedback on performance to improve and accelerate competency acquisition as opposed to summarizing the achievement status of a student [8,9]. Any learning activity has potential value as formative assessment from oral discourse to conventional quizzes [10]. Three core principles form the basis for formative assessment [11]. Firstly, formative assessment should be viewed as an integral part of instruction, and it should be used in real time for guiding learning process. The material provided to students should depend on their current state of knowledge

and understanding. Secondly, formative assessment fosters student involvement. Students are not punished for their mistakes since assessment does not affect the final grade. Instead they can use assessments for self-guidance and to monitor their progress towards learning objectives. Thirdly, formative assessment requires constructive feedback. Feedback is the key element in formative assessment [12], it is claimed to be the single most powerful factor in enhancing achievement [13]. Studies indicate that feedback may improve learning from about 0.4 Standard Deviation (SD) to 0.8 SD [14], however a critique of these results exists [2]. One of the goals of educational process is identifying the gaps between what is known and what is aimed to be known [15]. Feedback which is separated from competency demonstration by days or weeks, as it happens with paper based tests that have to be manually graded, have diminished value [10]. Feedback in a form of a mark as provided by traditional tests has a limited use to a student as it is not congruent with good feedback practices [16]. Simply put, a mark does not specify the deficiencies in the student's knowledge which they should address in the future [17]. Another strength of formative assessment is that it can aid in using students' strengths and weaknesses to frame learning goals and monitor progress towards them [18].

In addition to linear traditional tests with fixed test items, adaptive tests based on Item Response Theory (IRT) have been proposed [19] and implemented [20]. In IRT computer adaptive tests each test item is dynamically chosen based on the student's answers. This can improve the precision of testing by adapting to the knowledge demonstrated by each student and shorten test times [21]. These advantages come at cost of introducing several new issues. Most problematic of which are greater test creation complexity because of the necessity of calibration of question pool and the premise that different questions measure one common trait [22]. This premise is not always useful when assessing separate competencies even in the context of Multidimensional Item Response Theory (MIRT) [23], which can even further exacerbate the challenge of item calibration. In practice, the main drawbacks remain the complexity and time required to create a test, which may dissuade teachers from using it as a learning tool.

Another alternative to traditional methods is Elo rating based approach [24]. Elo rating system was originally developed for chess skill estimation and ranking. But since has been adapted for use in education [25,26]. Within the system each participant is assigned a numerical rating and in case of educational applications so are the test items. The expected probability that a player wins, or a student successfully completes the task is then given by the logistic function and depends on the ratings involved. In this formulation Elo rating system estimates the probability of correct answer in the same way as one parameter IRT model (Rasch model) [27]. What differs is the parameter estimation procedure. Analogically with MIRT multivariate Elo extensions have been tested [28]. Elo rating-based systems, like IRT systems, depend on knowledge of item difficulties to estimate students' proficiency.

Apart from IRT and Elo rating there are other approaches to model learning [29], such as performance factors analysis [30] and Bayesian knowledge tracing [31]. However, these models can be equally hard to implement and use due to non-trivial calibration and set up procedures.

In this paper, an alternative testing method based on Upper-Confidence Bound [32,33] Multi-Armed Bandit algorithm is proposed and tested. Multi-Armed Bandit (MAB) family of algorithms is named after a problem for a gambler who has to decide which arm of a K-slot machine to pull to maximize his total reward in a series of trials [34]. These algorithms capable of negotiating exploration–exploitation trade-offs are applied in several modern applications such as advertisement placement, website optimization, and packet routing [35]. There are emerging applications of MAB algorithms in education for optimal learning material selection [36–39].

The problem of choosing next question during a formative quiz can be modeled as a MAB problem. To the best of our knowledge no applications of MAB algorithms for formative assessment item selection have been proposed. Research supports that the key element of formative assessment is feedback, awareness of gaps between current students' knowledge and their aims, and where to go next to alleviate those deficiencies. The proposed in this paper assessment method addresses this need by quickly identifying the lacking areas of knowledge and thoroughly exploring them in order to assist

further learning. The process can be viewed through the lens of J. Hatties three question feedback model [15]. To utilize the algorithm the teacher must first form topics or competences (Where am I going?). The algorithm quickly identifies lacking areas of knowledge (How am I going?) and explores the topic in detail helping further instruction (Where to next?). This approach in combination with presently widespread mobile devices has the potential to mitigate the aforementioned issues such as test creation complexity and long test times, while providing accurate formative assessment data compatible with J. Hattie three question feedback, Competency Based Learning and Assessment methodologies.

2. Materials and Methods

2.1. Modelling Assessment as UCB Problem

When tutoring, a teacher will often engage in a dialogue with a student. The teacher may ask the student a series of formative questions in order to diagnose the gaps in student's knowledge. Assume the material consists of two topics, and the teacher asked 5 questions on each topic. The knowledge about first topic appears to be in a worse shape, two incorrect answers, than the knowledge on the second topic, one incorrect answer (see Table 1).

Table 1. Rudimentary two topic quiz model.

Question	1	2	3	4	5	6	7	8	9	10
Topic 1	+		+	−		+				−
Topic 2		+			+		+	−	+	

If there is time for five more questions before proceeding with didactic instruction, the teacher must face the dilemma of which topic should they explore with further questions? If the two incorrect answers to questions on first topic are attributable to bad luck due to small sample size, should the teacher explore first or second topic? What if there are more than two topics?

The family of bandit algorithms are designed to cope with uncertainty by balancing exploration and exploitation [40]. However, when applied to formative assessment the exploitation component is non obvious, as ultimately the goal is to explore the knowledge of the student. The algorithm should probe and explore the different topics and engage in focused questioning, exploiting those which are possibly in most need of instruction. This presents an opaque bandit problem where a unique answer, reward, is observed at each round, in contrast with the transparent one where all rewards are observed [34]. Thus, in context of assessment, a sequential allocation problem is obtained when the assessor has to choose from many questions from multiple topics, bandits, and has to repeatedly choose a topic to explore, which bandit arm to pull. When choosing next question to ask the decision should depend on the history of already known answers. Then a policy is the mapping from the individual history of the student to actions (questions to be asked of the student).

Suppose student's knowledge on number of topics $T = \{1, 2, \ldots, k\}$. The reward in case of a multiple-choice quiz with either correct on incorrect answer to each question $X_r \in \{0, 1\}$ is binary valued. Each topic corresponds to an unknown probability distribution. There exists a vector $\mu \in [0, 1]^k$ such that the probability that $X_r = 0$ given the algorithm chose topic $T_r = t$ is μ_t. This kind of environment is called a stochastic Bernoulli bandit. If the mean vector associated with the environment was known, the optimal policy is to always choose a question on one topic $t^* = \text{argmin}_{t \in T} \mu_t$. This will result in the exploration of the weakest area of student's knowledge, so as to aid in the further instruction. The regret over the n questions is

$$R_n = E\left[\sum_{r=1}^{n} X_r\right] - n \min_{t \in T} \mu_t \tag{1}$$

where the expectation E is with respect to stochastic environment and policy. However, in practical setting, the number of questions on one topic is usually rather limited due to the scope of the curriculum

and the question pool. As is the length, or the horizon, of the quiz. Thus the value of calculated this way regret is of little practical value.

The main challenge of the task is finding the weakest topic of a student. To do so the algorithm must explore different topics and exploit particular topic to obtain more accurate estimation of the student's level of knowledge on that subject. This basic exploration–exploitation dilemma is the key to obtaining a good strategy. A heuristic principle for dealing with this issue chosen in this paper is optimism in face of uncertainty and an algorithm which operates on this principle Upper-Confidence Bound (UCB). UCB algorithm is one of the simplest algorithms that offers sub-linear regret. The algorithm suggest choosing the action with the largest upper confidence bound, or in case of our model a topic with the smaller lower confidence bound. Then the question number n chosen on a topic t will be

$$t_n = \underset{t \in T}{argmin}\left(\mu_t - C\sqrt{\frac{logn}{N_t}}\right) \tag{2}$$

where C is a constant that can be chosen to regulate the impact the second exploration component has on the choice of the topic, and N_t is the number questions on the topic has been asked so far. As the number of questions on the topic increases, so the uncertainty and the exploration term of the formula decrease [40]. Thus the algorithm will seek out the weakest topics of knowledge for a student, once identified it will thoroughly question the student on said topics. This is pedagogically valuable because, once identified the lacking topic knowledge can be corrected. In addition, the algorithm will gather a more fine-grained information on the weakest topic by "exploiting it", which will be useful in post assessment knowledge correction. In the case when the item pool of the topic is exhausted the algorithm chooses the topic with the second smallest value as estimated by Formula (2).

2.2. Simulated Students' Experiments

A number of experiments with simulated students were carried out before proceeding with testing using empirical data. Number of simulated students for each experiment was 1000, unless stated otherwise. Each simulated quiz had a number of question on two or more topics. Each topic, then, would be represented as a vector of weights for each question in the quiz. Each weight would represent the relevancy of the question to the topic. In this paper only experiments with binary weights, 0 or 1, where carried out. Moreover, each question was assumed to belong only to one topic. The number of question items on each topic were set to be equal. Each simulated student had a vector equal in length to the number of the questions in the quiz, where each element represented the knowledge on one question, either yes or no. For all simulation experiments, unless stated otherwise, the inter-topic correlation of answers was random. The probability that the student will know the answer on a topic p_t had uniformly distributed random bias between 0 or 1.

2.3. Real Students' Assessment Methodology

In this study, feasibility of application of UCB algorithm to formative assessment is explored using the data collected from a 60-question quiz covering 15 subtopics. (i.e., network topologies, networking devices, Internet Protocol version 4, Ethernet, cloud computing services). The assessment was held at Vilnius Gediminas Technical University, Lithuania (on 25 April 2019). The test length and item pool size were set to 60 questions to keep quiz length close to an hour. Number of topics was set to 15 because it is the number of lectures in the course. The quiz was designed to assess students' knowledge of basic computer networking and cloud computing technologies. In total 104 undergraduate, sophomore and junior (third year), students from 7 different groups where tested. All of the students took the test at the same place and time, (Saulėtekio al. 11, Vilnius, from 12:30 to 13:30). The question pool contained questions of varying difficult (hardest question was answered by 16 students, easiest by 103). This was done to test robustness of the algorithm, as it is meant as an alternative to methods

that require item calibration. Therefore the algorithm must be able to work with items of varying and priory unknown difficulty.

All questions in the 60-question quiz where multiple choice questions with four options, only one of which was correct. Questions on the same topic were made to never appear in consequence so as to lessen the impact of deductive reasoning over the knowledge of the subject. Students were not allowed to assist each other during the assessment. All students were also informed that if they so desire the test will have no summative impact on their grade and will serve exclusively formative function. Every student gave a written permission allowing their anonymized data to be used for scientific research. The answers to questions have been aggregated in a comma separated file (csv), anonymized and later processed using python software written for the purpose of this experiment.

The true knowledge of a student on each topic, the ground truth, was calculated by dividing the number of correct answers within a topic by a total number of questions within that topic. In the experiment with participation of real students the number of correct answers was known because all students were required to answer all questions in the item pool. After complete knowledge of the test material for each student was known, the algorithms would question the database. The accuracy was then measured as a relationship between an estimated student knowledge from the incomplete information accessed by the algorithm and the complete information in the database. Formulas used for accuracy calculation are provided in the following statistical analysis section.

2.4. Statistical Analysis

The accuracy (performance) of the test was established based on Positive Predictive Value (PPV), which defines the probability of supplying the correct learning material to a student after the formative assessment and evaluated according to the formula, $PPV = TP/(TP + FP)$ for one student. Where TP is True Positive, or the number of correctly identified weakest topics for a student. The number of weakest topics is not always one, because topic proficiency is assumed to be equal to an expected value of an answer on a topic question, a Bernoulli variable $E(T) = p_t$. This value can be the same for several topics, in that case it is assumed that the student would equally benefit from instruction on any of the topics. The FP, False Positive, is a number of incorrectly identified topics for which topic mean μ_t is larger than the smallest mean, μ_m. For the group of students average of individual accuracies was taken.

Where applicable experimental results where expressed as mean ± Standard Error of the Mean (SEM). Correlation matrix of questions for heat-map visualization was computed using Pearson correlation coefficient using Pandas Python data analysis library.

Variance of answers on questions on one topic in simulation experiments was calculated using $\mathrm{Var}[T] = p_t(1 - p_t)$ formula for Bernoulli distribution. The probability of correct answer on the topic p_t was known and controlled for each topic T to observe its effect on assessment accuracy. In the experiments where real students' variance was computed using same formula, p_t was estimated using formula

$$p_t = \frac{\sum_{s=1}^{n_s} \sum_{q=1}^{n_{qt}} a_{sq}}{n_s n_{qt}} \tag{3}$$

where s is a student, q is a question within a topic and a_{sq} is an answer of a particular student on a particular question within a topic, n_s and n_{st} are the number of students and questions within a topic, respectively.

3. Results

3.1. Impact of Exploration Constant on Accuracy

We start by presenting a set of simulations to systematically explore different properties of formative assessment using UCB algorithm. UCB algorithm efficiency is dependent on the constant C which regulates the impact of exploration term on the topic choice as can be seen in Formula (2).

To analyze this impact and to choose the most suitable C for assessing real students a number of experiments with synthetic students were carried out.

A cursory result for the C impact on assessment accuracy can be seen in Figure 1. UCB algorithm shows better performance for every plotted constant over randomly asked questions. Note that algorithm serving random questions never asks same question twice of the same student, thus it achieves 100% accuracy after serving all 64 question items. It is clear that exploration can have both positive and negative impact on accuracy as seen from better performance of $C = 0.45$ over $C = 0$ and $C = 1$.

From Figure 1 it is clear that UCB algorithm when applied to formative assessment has a potential to significantly shorten test length. With larger constant the algorithm displayed relatively bad accuracy at quiz lengths from about 15 to 30 questions. This can be explained by failure to exploit known bad topics in order to further explore topics about which little data is known. Finally, as seen from the best performance of $C = 0.45$ exploration component does have a positive impact on assessment accuracy.

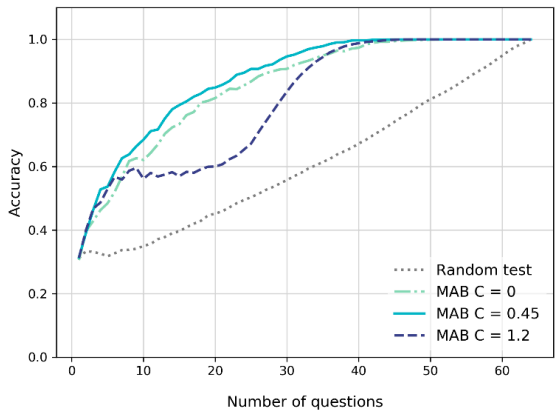

Figure 1. Assessment accuracy as function of number of questions in a quiz. A quiz contained a pool of 64 questions on 16 different topics and shows average accuracy for a class of 1000 synthetic students.

To choose the appropriate exploration constant for the real quiz, which had 60 questions (to set its duration at about 60 min), a following experiment with synthetic students was carried out. The number of questions was set to 64 in order to observe the importance of exploration in realistic scenarios: 4, 8, and 16 topics (see Figure 2). In the experiment we measured the minimal quiz length (number of questions) required to achieve accuracy greater than 95% in the class of 1000 synthetic students. The experiment was performed for every constant value from 0 to 2, with a step of 0.1 and the results were plotted in Figure 2. A conclusion we can draw from Figure 2 is that for every practical topic number in a 64-question test constant can be set to 0.5 for optimal results.

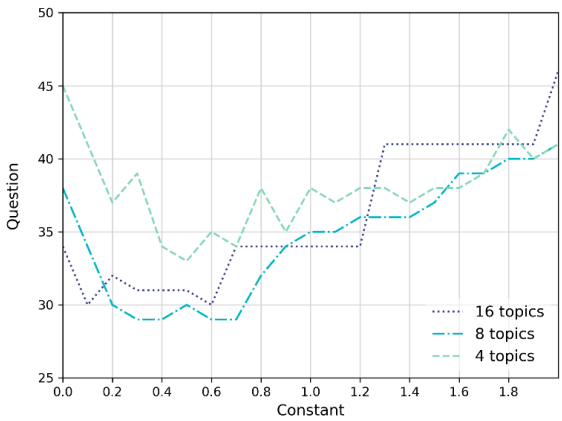

Figure 2. Number of questions needed to achieve 95% accuracy as a function of exploration constant C for 4, 8, and 16 topics and 64 questions.

3.2. UCB Advantage for Different Quiz Length

A more thorough exploration of quiz space has been performed to gauge the practicality of implementing UCB algorithm in different classroom situations. From Table 2 it is clear that UCB testing method becomes more valuable as pool of questions and number of topics increase. Number of simulated students for this experiment was 500.

Table 2. Reduction in quiz length necessary to achieve 95% accuracy conferred by using UCB algorithm with optimal exploration constant over traditional random question approach. All numbers are given for the optimal C value, where T—number of topics; Q—total number of questions. Green color indicates large reduction in quiz length, and red indicates lesser benefit.

Q \ T	4	8	16	32	64	128	256	512
2	33%	0%	8%	0%	24%	−2%	22%	24%
4	0%	29%	31%	35%	36%	44%	43%	52%
8		0%	40%	45%	50%	54%	58%	60%
16			0%	57%	50%	59%	67%	71%
32				0%	65%	52%	64%	73%
64					0%	62%	56%	66%

3.3. Assessment of Real Students

In Figure 3 are presented the results from the experiment with real students (n = 104). A formative assessment quiz included items from 15 topics, with 4 question items in each topic, for 60 total questions. Results show that in this scenario using traditional testing methods quiz length could be shortened from 60 questions to 55 questions if the goal was 95% accuracy in weakest topic identification. With use of UCB adaptive assessment, however, quiz length could be almost halved to 32 questions. Empirical results (orange and blue plots) are in line with projections drawn from simulations (grey plot). For this particular quiz and group of students UCB would offer a reduction in quiz length by 23 questions if we aim for same (>95%) accuracy.

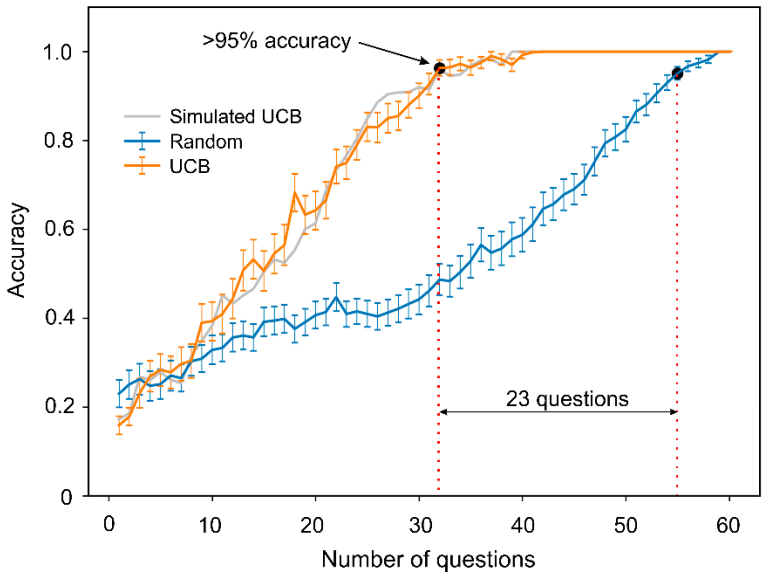

Figure 3. Average accuracy of formative assessment for 104 students. Simulated Upper-Confidence Bound (UCB) line is for 104 synthetic students to show that synthetic experiments are congruent with reality. The error bars represent SEM.

The ease at which the weakest topic of a student can be identified is dependent on how strongly the answers on the same topic are correlated. At answer variance equal to zero, it is sufficient to ask only one question to know the student's knowledge of the rest of the items within the topic.

However, when answers are uncorrelated, estimation must be harder, and might defeat entire premise of the proposed UCB testing model. Thus an experiment to measure the effect of knowledge correlation on items within one topic on method effectiveness was carried out. The impact of answer variance on testing accuracy can be seen in Figure 4.

As anticipated, identifying weak topics is trivial for unrealistically strongly correlated answers. However, even for uncorrelated answers UCB performs twice as good as random questioning. Also plotted in the Figure 4 are variances calculated from answers of real students, $N = 104$ (15 topic 60 question quiz).

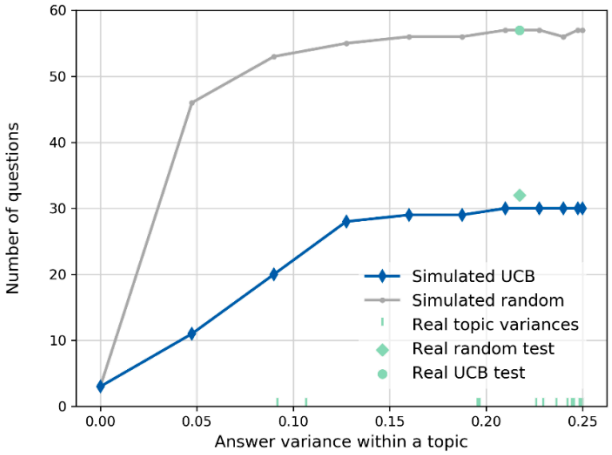

Figure 4. The impact of answer variance within topics on testing accuracy (number of questions needed to achieve 95% accuracy in a 60 question 15 topic quiz). The green markers indicate variance calculated using data from real tests, square and circle position along x axis represents average variance.

Correlation matrix for answers between different questions has been constructed and is shown in Figure 5. It indicates relatively low answer correlation even for inter topic questions. Answers on each topic where grouped together for this illustration (i.e., answers 1, 2, 3, 4 all belong to same topic). The matrix in accordance with Figure 4 shows low general correlation of answers within one topic especially for items 20–24 (questions about the physical layer of OSI model).

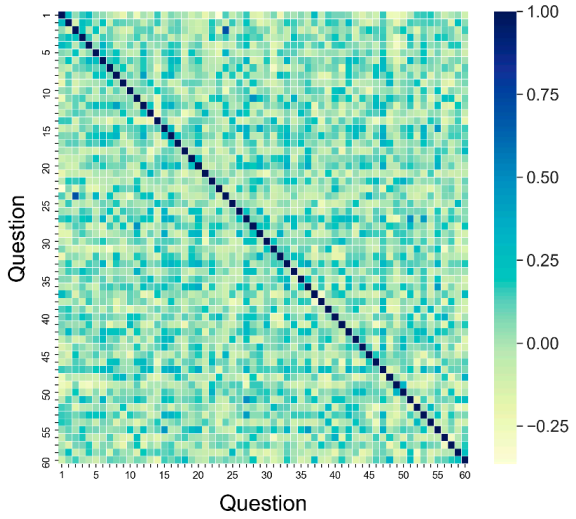

Figure 5. Correlation matrix (Pearson coefficient) for quiz answers.

To assess shorter quiz lengths and algorithm behavior at 95% accuracy value a more detailed look at accuracy distribution is provided in Figure 6. As seen from the figure for 60 question, 15 topic quiz the algorithm rarely displays accuracies between 50% and 99% for individual students.

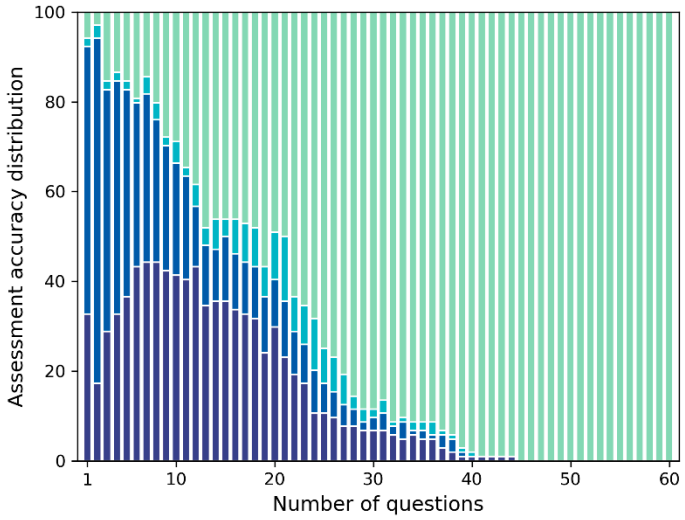

Figure 6. Distribution of assessment accuracies for students, where dark blue <1%, blue 1–50%, cyan 50–99%, and green >99% accuracy.

At quiz lengths between 10 and 32 questions a substantial portion (4–38%) of students were assessed very poorly, with less than 1% accuracy. Similarly, even at >95% accuracies, there can be a minority of students with incorrect weakest topic estimates. The root causes may be small correlation of answers (Figures 4 and 5) in the sample used for testing and small number of items within a topic (only 4). There was an increase in the fraction of students with wildly wrong estimates (accuracy <1%) up to question 7. This is because initial estimated knowledge on all topics is set to 0.5 in our implementation of the algorithm. The algorithm shows a steady increase in accurately (>99%) diagnosed students with no abnormalities.

4. Discussion

UCB algorithm has a potential to significantly reduce assessment length without the loss of accuracy. Even for very short tests with few topics the algorithm offers significant reduction of test length (Table 2). However, there is no advantage for quizzes in which each topic contains only one question, at which point the notion of topic loses its pedagogical meaning. There is evidence that time allocated to study positively affects student performance [41], thus reducing the time spend on assessment is desired. The experiments with real students support the conclusions drawn from the simulations (Figure 3). Such strong change in quiz length has the potential to change the dynamics in the classroom, because the instructor would not need to spend an entire lesson just for formative assessment. This in turn, may increase opportunities to provide personalized feedback to students linked to better performance [14,42]. This advantage comes at no cost in quiz creation complexity, unlike IRT and Elo rating based systems where quiz creation can be prohibitively complex in some situations due to item calibration problem [22,27]. Because of this fundamental difference we do not compare performance of UCB to IRT and Elo based algorithms. Such comparison would not discredit either approach: If UCB performs worse (as is safe to assume), its simplicity of use and independence from question item calibration makes it an interesting alternative to traditional assessment.

It is clear that exploration component of the algorithm becomes more important with the increase in number of items within a topic. This can be observed in bad performance of algorithm in 4 topic quiz when C was set to 0 in Figure 2. It took 45 questions to reach >95% assessment accuracy.

Also, unintuitively, it takes less questions to identify weakest topics of students' when there are more topics when the size of the item pool is kept constant.

Compared to IRT and Elo algorithms UCB will obtain less information about the student in a mathematical sense [27,43], and this can be seen as a disadvantage. However, not all information is equally pedagogically valuable [12,16]. For example, assume we are assessing student's knowledge on two topics. We are nearing the end of the quiz and have determined that knowledge on first topic is adequate, but lacking on the second topic. Because of the nature of UCB algorithm topic two is more explored than topic one, therefore we expect to gain less information by exploring second topic. However, from pedagogical point of view the information on topic two can be more valuable to address the assessment needs according to the three question feedback model [15]. According to the assessment, topic two is in need of instruction. If we are going to proceed to teach it, we can use the extra diagnostic data to save time and effort by not teaching what the student already knows. Meanwhile we have no immediate use for the more precise data about first topic.

Simulation data presented in Figure 4 and empirical results (Figures 3 and 5) suggest that the UCB assessment approach can offer significant reduction in quiz length for any practical item pool size and inter-topic variance of answers. This is an important result because it indicates suitability of the method for any grouping of questions regardless of how correlated the answers are within one topic. Method effectiveness stays almost constant for any observed answer variance, which makes it easy to predict quiz length. This allows a teacher to group questions on each topic as they see fit according with syllabus and the learning material at hand, regardless of existence or lack of a common latent trait underlying the items within a topic. This separates UCB method from IRT and Elo based alternatives which depend on the assumption of common latent trait [22,28].

At quiz lengths between 10 and 32 questions a substantial portion (4–38%) of students were assessed very poorly implies that using shorter assessments is morally questionable, as the majority of the students will receive a very accurate guidance, while the rest will be tutored on topics which they already know. This presents a problem for more important formative assessments (i.e., entire semester assessment). This property of the algorithm can be offset my small increase in quiz length as seen in Figure 6.

5. Conclusions

Presented in this paper novel approach to formative assessment based on UCB algorithm shows promising results when compared to traditional assessment methods. This approach can significantly reduce quiz length without reduction in accuracy. For quizzes with item pool equal to 8 questions the reduction is 29%, for quizzes with item pool of 512 question it is 73%. Variance of answers to questions within same topic has little impact on assessment accuracy for empirically observed values (0.1 to 0.25), thus the algorithm is suited for situations where items do not necessary measure same latent skill or trait. However, distribution of student accuracies within a class is non-normal. Even at high average class accuracies (95%), the majority of accurately assessed students is offset by a small minority of students for whom weakest topics where incorrectly identified. To offset this property of the algorithm we recommend that educators target >99% accuracy for course-crucial UCB formative assessments. We believe UCB based formative assessment has pedagogical potential for practical applications and should be further explored. Unlike IRT and Elo rating-based assessment methods UCB based assessment requires no question item calibration and does not depend on the debatable premise that different questions measure same latent trait. As consequence UCB method belongs to a different, sparsely explored class of easy to implement and maintain formative assessment solutions. It may prove to be a fresh and viable alternative to traditional linear assessment in situations where IRT and Elo methods were deemed too complex to implement and maintain. In the future a comparative study of UCB assessment method with established item calibration dependent methods may be of interest.

Author Contributions: Conceptualization, J.M., V.N.; methodology, J.M., and V.N.; writing—original draft preparation, J.M.; writing—review and editing, J.M., and V.N.; Supervision, J.M.

Funding: This research received no external funding.

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

References

1. Black, P.; Wiliam, D. Assessment and Classroom Learning. *Assess. Educ. Princ. Policy Pract.* **1998**, *5*, 7–74. [CrossRef]
2. Bennett, R.E. Formative assessment: A critical review. *Assess. Educ. Princ. Policy Pract.* **2011**, *18*, 5–25. [CrossRef]
3. 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; pp. 1–11.
4. 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**. [CrossRef]
5. Scriven, M. *The Methodology of Evaluation*; Social Science Education Consortium; Purdue University: West Lafayette, Indiana, 1967.
6. Bloom, B.S. Some theoretical issues relating to educational evaluation. In *Educational Evaluation: New Roles, New Means: The 63rd Yearbook of the National Society for the Study of Education Part II*; University of Chicago Press: Chicago, IL, USA, 1969; pp. 26–50.
7. Bloom, B.S. Mastery learning: Theory and practice. In *Mastery Leazrning*; Holt Rinehart & Winston: New York, NY, USA, 1971.
8. Sadler, D.R. Formative Assessment: Revisiting the territory. *Assess. Educ. Princ. Policy Pract.* **2007**, *5*, 77–84. [CrossRef]
9. Moss, C.M.; Brookhart, S.M. *Advancing Formative Assessment in Every Classroom: A Guide for Instructional Leaders*; ASCD: Alexandria, VA, USA, 2019; ISBN 1416626727.
10. Gareis, C.R. Reclaiming an important teacher competency: The lost art of formative assessment. *J. Pers. Eval. Educ.* **2007**, *20*, 17–20. [CrossRef]
11. Group, A.R. *Assessment for Learning: Beyond the Black Box*; Qualifications and Curriculum Authority: Coventry, UK, 1999; ISBN 0856030422.
12. Sadley, D.R. Formative assessment and the design of instructional systems. *Instr. Sci.* **1989**, *18*, 119–144. [CrossRef]
13. Marzano, R.; Pickering, D.; Pollock, J. *Classroom Instruction that Works: Research-Based Strategies for Increasing Student Achievement*; Pearson: London, UK, 2001; ISBN 0871207338.
14. Shute, V.J. Focus on formative feedback. *Rev. Educ. Res.* **2008**, *78*, 153–189. [CrossRef]
15. Hattie, J.; Timperley, H. The Power of Feedback. *Rev. Educ. Res.* **2007**, *77*, 81–112. [CrossRef]
16. Nicol, D.; Macfarlane-Dick, D. *Rethinking Formative Assessment in HE: A Theoretical Model and Seven Principles of Good Feedback Practice*; Higher Education Academy: London, UK, 2006.
17. Guskey, T.R.; Bailey, J.M. *Developing Grading and Reporting Systems for Student Learning*; Corwin Press: Thousand Oaks, CA, USA, 2001; ISBN 080396854X.
18. 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] [PubMed]
19. Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*; Routledge: London, UK, 1980; ISBN 1136557245.
20. Huang, Y.M.; Lin, Y.T.; Cheng, S.C. An adaptive testing system for supporting versatile educational assessment. *Comput. Educ.* **2009**, *52*, 53–67. [CrossRef]
21. 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]
22. Wainer, H. *Computer-Adaptive Testing: A Primer*; Lang. Learn. Technol.; Erlbaum: New York, NY, USA, 2001.
23. Reckase, M.D. The past and future of multidimensional item response theory. *Appl. Psychol. Meas.* **1997**, *21*, 25–36. [CrossRef]
24. Elo, A.E. *The Rating Of Chess Players, Past & Present*; Ishi Press: Philadelphia, PA, USA, 1978; ISBN 9780923891275.

25. Wauters, K.; Desmet, P.; Noortgate, W. Monitoring learners' proficiency: Weight adaptation in the Elo rating system. In Proceedings of the EDM 2011—Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, The Netherlands, 6–8 July 2011.
26. Brinkhuis, M.; Maris, G. *Dynamic Parameter Estimation in Student Monitoring Systems*; Measurement and Research Department Reports; CITO: Arnhem, The Netherlands, 2009.
27. Pelánek, R. Applications of the Elo rating system in adaptive educational systems. *Comput. Educ.* **2016**, *98*, 169–179. [CrossRef]
28. Doebler, P.; Alavash, M.; Giessing, C. Adaptive experiments with a multivariate Elo-type algorithm. *Behav. Res. Methods* **2015**, *47*, 384–394. [CrossRef] [PubMed]
29. Desmarais, M.C.; Baker, R.S.J.D. A review of recent advances in learner and skill modeling in intelligent learning environments. In *User Modeling and User-Adapted Interactiont*; Springer: Berlin/Heidelberg, Germany, 2012.
30. Pavlik, P.I.; Cen, H.; Koedinger, K.R. Performance factors analysis—A new alternative to knowledge tracing. In *Proceedings of the Frontiers in Artificial Intelligence and Applications*; IOS Press: Amsterdam, The Netherlands, 2009.
31. Corbett, A.T.; Anderson, J.R. Knowledge tracing: Modeling the acquisition of procedural knowledge. In *User Modeling and User-Adapted Interactiont*; Springer: Berlin/Heidelberg, Germany, 1994.
32. Lai, T.; Robbins, H. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **1985**, *6*, 4–22. [CrossRef]
33. 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]
34. Vermorel, J.; Mohri, M. Multi-armed bandit algorithms and empirical evaluation. In Proceedings of the Lecture Notes in Computer Science Oun (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Porto, Portugal, 3–7 May 2005.
35. Bubeck, S.; Cesa-Bianchi, N. *Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems*; Foundations and Trends®In Machine Learning: Hanover, MA, USA, 2012; Volume 5, pp. 1–122.
36. Clement, B.; Roy, D.; Oudeyer, P.-Y.; Lopes, M. Multi-armed bandits for intelligent tutoring systems. *arXiv* **2013**, arXiv:1310.3174.
37. Clement, B.; Roy, D.; Oudeyer, P.-Y.; Lopes, M. Online optimization of teaching sequences with multi-armed bandits. In Proceedings of the 7th International Conference on Educational Data Mining, London, UK, 4–7 July 2014.
38. 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 to Cite This Version. In Proceedings of the 7th International Conference on Educational Data Mining, London, UK, 4–7 July 2014.
39. 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, Raleigh, NC, USA, 29 June–2 July 2016.
40. Szepesvari, C.; Lattimore, T. *Bandit Algorithms*; Foundations and Trends®In Machine Learning: Hanover, MA, USA, 2019; ISBN 9781449341336.
41. Bratti, M.; Staffolani, S. Student Time Allocation and Educational Production Functions. *Ann. Econ. Stat.* **2013**, *111–112*, 103–140. [CrossRef]
42. Marzano, R.J.; Gaddy, B.B.; Dean, C. *What Works in Classroom Instruction*; Classr. Instr. That Work; Mid-Continent Research for Education and Learning: Denver, CO, USA, 2000.
43. Huang, C.J.; Chen, H.X.; Chen, C.H. Developing argumentation processing agents for computer-supported collaborative learning. *Expert Syst. Appl.* **2009**, *36*, 2615–2624. [CrossRef]