Evaluating Bayesian Knowledge Tracing for Estimating Learner Proficiency and Guiding Learner Behavior

Shreyansh Bhatt
Amazon.com
Seattle, USA
bhattshr@amazon.com

Jinjin Zhao
Amazon.com
Seattle, USA
jinjzhao@amazon.com

Candace Thille
Amazon.com
Seattle, USA
cthille@amazon.com

Dawn Zimmaro
Amazon.com
Austin, USA
dzimmaro@amazon.com

Neelesh Gattani
Amazon.com
Seattle, USA
neeleshg@amazon.com

ABSTRACT

Open navigation online learning systems allow learners to choose the next learning activity. These systems can be instrumented to provide learners with feedback to help them choose the next learning activity. One type of feedback is providing an estimate of the learner's current skill proficiency. A learner can then choose to skip the remaining learning activities for that skill after achieving proficiency in that skill. In this paper, we investigate whether predicting proficiency and communicating it to learners can save time for learners within a course. We evaluate the accuracy of the BKT based proficiency prediction framework for learner's proficiency prediction which considers one attempt per question. We extend the proficiency prediction framework to include multiple attempts at individual questions and show that it is more accurate in proficiency prediction than BKT based proficiency prediction framework. We discuss the potential implications of attempt enhanced framework on the learners' behavior for open navigation online learning systems.

Author Keywords

Knowledge tracing; Baysian Knowledge Tracing; Mastery prediction; Learner behavior analysis

CCS Concepts

•**Applied computing** → **Learning management systems; Interactive learning environments;** •**Human-centered computing** → *User models; User studies;* •**Mathematics of computing** → **Bayesian networks;**

INTRODUCTION

Online learning systems with open navigation offer courses/learning experiences to learners and provide feed-

back to learners to navigate to learning activities depending on the skills they want to develop, rather than the system automatically selecting the next learning activity. Such a system provides feedback to a learner after he/she goes through a content block and attempts a question associated with the content block. These questions are tagged with skills that can be used to predict the learners' proficiency in the associated skills. Skill proficiency feedback can help a learner to choose the next learning activity. For example, after achieving a proficiency, learner can choose to skip the rest of the practice opportunities for that particular skill.

BKT models learner proficiency, often called mastery, as a binary variable [2]. Most online learning systems, such as intelligent tutoring systems, use BKT to select the next learning activity to present to the learner. Most of the studies evaluated BKT and its extensions such as inclusion of forget parameter [3], item difficulty [5], learner based individualization [11], and time between attempts [7] for predicting the probability of correctly answering the next question. Recently, deep learning based knowledge tracing (DKT) achieved state-of-the-art performance in predicting the probability of correctly answering the next question [6]. Several extensions of DKT and memory augmented variation reported better next attempt's outcome prediction accuracy [10][8][9][12]. These approaches require a large number of training records and have shown to work well in recommending the next learning activity. Evaluating proficiency predicted by BKT is different from predicting the probability of correctly answering the next question because a learner can answer the next question correctly from either a proficient state or a non-proficient state. As we communicate to learners about their learning, we want to evaluate the impact of feedback about proficiency instead of the answer to the next question.

In this paper, we evaluate a BKT based proficiency prediction framework. The framework consists of predicting the probability of being in the proficient state using the BKT model after the learner makes a minimum number of attempts for a skill. The minimum number of attempts are determined either

empirically or based on the behavior of the other learners in the same or a related course.

We used three courses designed for adult workforce skill development and offered through an open navigation based online learning system to evaluate the proficiency prediction framework. We created the ground truth proficiency data for this purpose. We created ground truth proficiency labels with inputs from human experts for two courses and from learner for one course.

The open navigation learning system is designed for adult workforce skill development. The system offers learning experiences to a diverse set of learners. Learners can choose to focus on appropriate skills, depending on the proficiency feedback, to maximize learning for the amount of time spent on the learning experience. To measure the effect of communicating proficiency to learners and the effect of it on learners' behavior, we compute the average time saved per learner. We also analyze whether explicitly communicating the proficiency to learners can influence their behavior (and save time).

The online learning system provides a hint if a learner's attempt to a question is wrong. Hence, each attempt is a learning opportunity as the learner receives targeted feedback for an incorrect attempt. To consider this, consistent with the attempt enhanced BKT proposed by Pardos et al. [4], we extend the BKT framework to include multiple attempts per question in computing proficiency. As we consider each attempt as a learning opportunity, we also revise the minimum number of opportunities computation. We propose to compute the minimum number of opportunities required to achieve proficiency in terms of each attempt made to a skill and in terms of the total number of questions successfully answered by a learner.

We are interested in exploring the following research questions,

- RQ1: Can predicting and communicating proficiency save time for learners?

- RQ2: Can BKT accurately predict proficiency?

- RQ3: Can incorporating attempt in BKT and the minimum number of trials estimated as attempts and question improve proficiency prediction accuracy?

PROFICIENCY PREDICTION FRAMEWORK
In this section, we describe the BKT based proficiency framework and its attempt enhanced variant.

In our proficiency prediction framework, we use a knowledge tracing model (such as BKT) to compute proficiency probability. For BKT, it is the probability of a learner being in the proficient state given the model parameters and observations (learner's assessment outcomes). If a learner achieves at least 0.9 probability from the prediction model after attempting at least the minimum number of questions/assessments, then the learner is considered proficient. In our BKT based proficiency prediction, we compute a weighted score per question as w_q. $w_q = 1.0 - \frac{c_a}{t_c}$. Here, t_c is the total number of choices for the question and c_a is the total number of attempts it took the learner to answer the question correctly. All the attempts larger
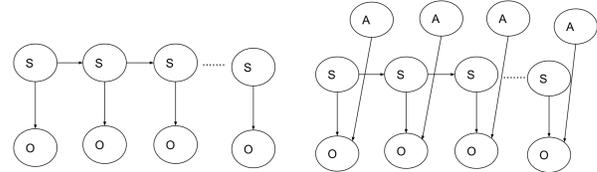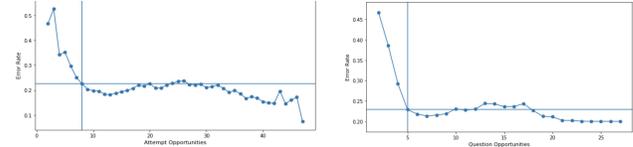

Figure 1: BKT and Attempt enhanced BKT.

Figure 2: Minimum number of opportunity calculation.

than c_a are ignored. If a learner has the weighted score more than 0.5 then the learner is considered to have a successful response to that question. In this framework, the minimum number of opportunities is set to four questions. Next, we describe the BKT model and its attempt enhanced variant.

BKT is a two-state Hidden Markov Model where the proficient and not-proficient states are hidden states and observable states include 0 or 1 (indicating the correctness of the learner's answer to the question). Using four parameters, *p_slip, p_guess, p_trans, p_init*, BKT computes the probability of being in the proficient state given the observed sequence for the learner [2].

We incorporate attempt in the knowledge tracing model as identified by Pardos et al. [5][4]. Figure 1 shows the BKT model diagram on the left and the attempt enhanced BKT model on the right. In the attempt enhanced BKT, observation depends on the proficient state as well as learner's attempt for a question. The attempt node is a multinomial node which takes the value up to four as we allowed up to four attempts for each question. Any attempt after the fourth attempt was not considered. We individualize *p_slip, p_trans, p_guess* parameters for each attempt. Slip and guess probability can be different depending on the attempt hence we individualize those two parameters. We also individualize the transition parameter to let learner transition to the proficient state with probability depending on the attempt number. As attempt is fully observable, the model can use appropriate *p_slip, p_guess, p_trans* depending on whether the attempt is the first, second, third, or fourth to any question.

Minimum number of attempts estimation
For the embedded assessment based learning system, a learning activity includes an expository content block followed by the assessment for the associated content.

As we consider multiple attempts to the same question in the attempt enhanced BKT, we compute the minimum number of opportunities for each skill as 1) Minimum number of attempts and 2) Minimum number of questions. Minimum number of attempts consists of all the attempts whether it is to the same question or a different question (associated with a particular skill). As a learner receives feedback for an unsuccessful

attempt, we consider that as a learning opportunity. The minimum number of questions indicates the number of questions a learner needs to answer to achieve proficiency in a skill.

We compute the minimum number of attempts (or minimum number of questions to answer) as the number of attempts at which 75% of learners achieved proficiency using learning curve [1]. A learning curve indicates error rate for every opportunity. This error rate for a given attempt 'a' is computed as $\frac{\sum_{i=1}^{N} 1.0 - p_a^i}{N}$. Here, p_a^i is a proficiency probability for the ith learner at the ath attempt. The first occurrence of learning curve crossing 0.25 error rate at the particular attempt is considered as the minimum number of attempts. To compute minimum number of questions, we use the same methodology except attempt number is incremented only after a learner attempts a new question. Note that the proficiency probability is still updated for multiple attempts. Figure 2 shows an example learning curve to find minimum number of attempts and minimum number of questions.

EXPERIMENTS AND RESULTS
Datasets, experiment setup, and results to follow.

Datasets & Proficiency Label Creation
We considered three courses offered through an online learning system for adult workforce learning experiences.

course 1: This course had 28 skills and 145 questions. 500 learners participated in this course and we collected proficiency information for 67 learners. After the course learners participated in a panel presentation and the panel of human experts evaluated these learners and provided proficient or non-proficient vote for these learners.

course 2: There were only three skills in this course. Skill 1 and skill 3 had five questions and skill 2 had four questions. 109 learners participated in this course. We had a domain expert provide the labels for each learner and skill as proficient or not proficient.

course 3: This course had 9 skills and 1500 learners participated in this course. Ground truth proficiency were computed based on learners' post-course evaluation that had questions related to each skill. Specifically, if the learner answered all the evaluation questions associated with the skill correctly, then we considered the learner as proficient.

Learner behavior analysis (RQ1)
Learners may decide to stop attending the skill, i.e., skip the content and practice question associated with the skill after the system communicates that they are proficient. On the other hand, learners may ignore the proficiency feedback if they think they still need to practice more for the given skill. For this, we compute the unattended questions by a learner after achieving proficiency for a skill. A learner has potentially saved total time it takes to answer all the unattended questions. Time to answer an unattended question is computed as the average time it took for the other learners (who answered this particular question) to answer the question.

Course	Time saved per learner
Course 1	74.81
Course 2	5.21
Course 3	24.83

Table 1: Average time saved per learner in minutes for the three courses. On an average, learners saved 75 minutes in course 1 and 25 minutes in course 3.

Metric	Version 1	Version 2
Attempts	0.76 ± 0.02	0.71 ± 0.02
Questions	0.46 ± 0.01	0.44 ± 0.01
Time per question	44.24 ± 0.5	43.04 ± 0.4
Time saved	4841 ± 80.6	4988 ± 85.7

Table 2: Version 2 recorded fewer average attempts, fewer questions, fewer time to answer question, and more time saved than version 1. P_value < 0.05 for t-test indicated that the difference was modest yet significant.

Table 1 shows the total time saved per learner for the three courses. Course 2 had a very few questions and a few learners hence the time saved wasn't significant. For course 1, we found that a learner on an average, saved about 75 minutes while on course 3, a learner on an average saved about 25 minutes by skipping questions associated with a skill after achieving the skill proficiency. As we have the job family and job level for each learner, we can also compute dollar amount saved.

Next, we show results for whether explicitly communicating proficiency can affect learners behavior. We conducted an experiment where two sets of learners were offered the same course in two different versions of the online learning system. In version 1, the learners had to access the proficiency information by navigating to the main course landing page separate from the course lessons that contain the learning activities and questions. In version 2, the proficiency information can be viewed on the pages on which the learners answer the questions. Moreover, the content is organized according to skill in version 2. Hence, learner can navigate to a specific section related to a skill as the learner views proficiency information.

Table 2 shows the results at the attempt level. For each question, we computed a ratio of the total number of attempts made to this question to the total number of users enrolled in the course. Table 2 shows the comparison at the question level. For this analysis, we computed the ratio of the number of users attempting a question to the total number of users enrolled in this course. We computed mean, standard error, and performed the t-test. For both attempt and question level analysis, we noticed that version 2 had a lower number of attempts and a lower number of questions attempted, on average. We also computed the average amount of time learners spent per question. On average, we noticed that learners spent 44.24 seconds per question in version 1 while 44.04 seconds in version 2 (p-value < 0.05). Hence, the time spent per question was slightly less in version 2 than version 1. We also noticed that the total time saved per learner was also increased in version 2 than version 1. Learners may be using proficiency data to skip questions in the course. Hence, it is critical to accurately compute the proficiency. For a course with 100 questions, it could save 30 minutes for a learner. A learner can spend this time on developing other skills in the course. As tens of

Dataset	Min Att	Min Questions
Course 1	4.0	3.0
Course 2	5.86 +- 1.27	4.93 +- 1.01
Course 3	8.7 +- 2.33	5.0 +- 1.22

Table 3: Mean and standard deviation for minimum number of attempts. For course 2, only one skill had non-zero slope learning curve leading to 0 standard deviation.

Dataset	BKT	Attempt Enhanced BKT
Course 1	0.8	0.91
Course 2	0.71	0.75
Course 3	0.89	0.89

Table 4: Attempt enhanced BKT with minimum attempt and question happens to be more accurate model for predicting the proficiency.

thousands of learners take this course, the system can save substantial amount of total learning time.

Attempt enhanced BKT and proficiency prediction (RQ2, RQ3)

We first show the minimum number of opportunity analysis as described in section 2.1. For each skill in the three courses, we computed minimum number of attempts and minimum number of questions. If the slope of the predicted learning curve is close to zero or does not decrease over the number of attempts, then such a skill was excluded from the analysis[1]. Table 3 shows the mean and standard deviation of the minimum number of attempts and minimum number of questions for the three courses. For the second course, we had only one skill that did not have close to zero or non-decreasing learning curve. Hence, the standard deviation is 0.

Next, we evaluated the BKT framework and its attempt enhanced variant using the ground truth proficiency data. For this, we considered a learner as proficient if the learner has at least "minimum number of attempts" or answers "minimum number of questions" and achieves at least 0.90 proficiency probability. We computed the proficiency probability using the attempt enhanced BKT. The model was trained (parameters were identified) based on the first 20% of the learners' records. Table 4 shows the F scores for the both models. BKT achieves adequate F-score (0.7+) in predicting the ground truth mastery. The attempt enhanced BKT achieves a better F score for two of the courses, especially for course 2 which has a small number of skills. For course 2, there are fewer questions designed per skill, and hence the learning opportunity provided by even incorrect attempts is important to consider.

We will continue evaluating the attempt enhanced BKT and its effect on learners. Four opportunities may be too few for some courses, and hence learners may continue to answer questions related to a skill even after the system communicates proficiency. The attempt enhanced model can provide better proficiency estimates so that learners can make more appropriate choices about how to spend their learning efforts.

CONCLUSIONS

We evaluated BKT for proficiency prediction using three courses offered through an open navigation online learning system. We found that BKT achieves adequate f-score (0.7+) in predicting proficiency. We noticed that proficiency prediction can save time to finish the course for learners. With a

modest effect, we notice that explicitly communicating proficiency to learners can influence learners' behavior and help them save time in learning. We also found that the attempt enhanced BKT is a promising approach to help learners make appropriate learning activity selection and can be more accurate in predicting proficiency than BKT. We will continue evaluating the attempt enhanced BKT to see whether it can help learners choose appropriate learning activities.

REFERENCES

[1] Olle Bälter, Dawn Zimmaro, and Candace Thille. 2018. Estimating the minimum number of opportunities needed for all students to achieve predicted mastery. *Smart Learning Environments* 5, 1 (2018), 15.

[2] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

[3] Mohammad Khajah, Robert V Lindsey, and Michael C Mozer. 2016. How deep is knowledge tracing? *arXiv preprint arXiv:1604.02416* (2016).

[4] Zachary A Pardos, Yoav Bergner, Daniel T Seaton, and David E Pritchard. 2013. Adapting Bayesian Knowledge Tracing to a Massive Open Online Course in edX. *EDM* 13 (2013), 137–144.

[5] Zachary A Pardos and Neil T Heffernan. 2011. KT-IDEM: Introducing item difficulty to the knowledge tracing model. In *International conference on user modeling, adaptation, and personalization*. Springer, 243–254.

[6] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *Advances in neural information processing systems*. 505–513.

[7] Yumeng Qiu, Yingmei Qi, Hanyuan Lu, Zachary A Pardos, and Neil T Heffernan. 2011. Does Time Matter? Modeling the Effect of Time with Bayesian Knowledge Tracing.. In *EDM*. 139–148.

[8] Yu Su, Qingwen Liu, Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Chris Ding, Si Wei, and Guoping Hu. 2018. Exercise-enhanced sequential modeling for student performance prediction. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[9] Zhiwei Wang, Xiaoqin Feng, Jiliang Tang, Gale Yan Huang, and Zitao Liu. 2019. Deep Knowledge Tracing with Side Information. In *International Conference on Artificial Intelligence in Education*. Springer, 303–308.

[10] Chun-Kit Yeung and Dit-Yan Yeung. 2018. Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. ACM, 5.

[11] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon. 2013. Individualized bayesian knowledge tracing models. In *International conference on artificial intelligence in education*. Springer, 171–180.

[12] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. 2017. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 765–774.