Effects of Algorithmic Transparency in Bayesian Knowledge Tracing on Trust and Perceived Accuracy

Kimberly Williamson
Cornell University
Ithaca, NY, USA
khw44@cornell.edu

René F. Kizilcec
Cornell University
Ithaca, NY, USA
kizilcec@cornell.edu

ABSTRACT

Knowledge tracing algorithms such as Bayesian Knowledge Tracing (BKT) can provide students and teachers with helpful information about their progress towards learning objectives. Despite the popularity of BKT in the research community, the algorithm is not widely adopted in educational practice. This may be due to skepticism from users and uncertainty over how to explain BKT to them to foster trust. We conducted a pre-registered 2x2 survey experiment (n=170) to investigate attitudes towards BKT and how they are affected by verbal and visual explanations of the algorithm. We find that ostensible learners prefer BKT over a simpler algorithm, rating BKT as more trustworthy, accurate, and sophisticated. Providing verbal and visual explanations of BKT improved confidence in the learning application, trust in BKT and its perceived accuracy. Findings suggest that people's acceptance of BKT may be higher than anticipated, especially when explanations are provided.

Keywords

Bayesian Knowledge Tracing, Data Visualization, Explainable AI

1. INTRODUCTION

Knowledge tracing can offer students and teachers a real-time understanding of what students have already learned and what they are still struggling with [7]. It provides actionable insights that can lead to better educational outcomes [16]. Among many types of knowledge tracing algorithms, Bayesian Knowledge Tracing (BKT) has been established and researched most extensively, as evidenced by the 114,000 Google Scholar results for "Bayesian Knowledge Tracing," 17,500 of which published since 2020. BKT has been tested to help students self-monitor their learning progress [4, 23], to help teachers understand what students have not learned yet [22], and to enable adaptive learning technologies that let students skip over the content they have mastered [18]. In contrast to the abundance of research on BKT, including hundreds of articles devoted to incremental enhancements of the original model [20], there are not many real-world applications that use BKT in practice. Some of the most widely used K-12 learning platforms like ASSISTments and Khan Academy decided against using BKT in favor of simpler models such as N-Consecutive Correct Responses (N-CCR) [13]. This raises questions about barriers to adopting knowledge tracing algorithms in educational practice. In particular, how much is the relative complexity and opacity of BKT responsible for its slow adoption? Platform providers may be concerned that educators and learners will not trust a model that cannot easily be explained to them [13, 12, 24, 25, 1].

The Technology Acceptance Model (TAM) posits that a user's acceptance and adoption of new technology is based on its perceived usefulness (PU) and perceived ease of use (PEOU) [9]. PU and PEOU are beliefs that can be influenced by external factors, such as providing additional information about a technology. According to TAM, learners' and educators' PU and PEOU are essential factors in the adoption of BKT in practice. Improving their perceptions could therefore increase the acceptance and adoption of BKT in real-world applications. Moreover, a better understanding of the mechanisms behind the acceptance of BKT is expected to inform the presentation of other knowledge tracing algorithms as well.

A large number of knowledge tracing algorithms have been developed over the years that could benefit from empirical evidence on how to explain them to users. Recent advances in artificial intelligence have inspired research into more complex algorithms such as deep knowledge tracing (DKT), which uses neural networks [17, 11]. With more complex algorithms that provide less insight into their inner workings, it becomes more important to understand how people's trust in the algorithm and its perceived accuracy might influence perceptions of usefulness and usability of a learning application [1]. Besides BKT and DKT, which are suitable for modeling understanding and sense-making, there are also logistic learning models, such as Additive Factor Models and Performance Factor Analysis [5, 19, 20], which model memory and fluency [20]. These two types of models can also be integrated into one [15]. While there are many types of models that can be examined, we choose BKT as an example knowledge tracing algorithms that is relatively simple and popular among researchers.

This research contributes causal evidence to address three important research questions. First, do people prefer to learn with BKT or N-CCR (N-Consecutive Correct Responses) in an ostensible high-stakes test scenario? Second, how is their preference related to specific attitudes, including their confidence in the learning system to do well on a test, their trust in the algorithm, and the perceived accuracy of the algorithm? And third, how do verbal and/or visual explanations affect people's attitudes and preferences over knowledge tracing algorithms? We answer these research questions with data collected from a pre-registered 2x2 factorial survey experiment.

2. BACKGROUND

One of the simplest knowledge tracing algorithms is N-CCR. It assesses student mastery by evaluating the number of consecutive correct responses for a particular skill. For example, the model determines that a student has learned fractions after correctly answering three fraction questions in a row. Although N-CCR is easy to understand, its simplicity can sometimes make it less accurate than BKT. Still, N-CCR has been used in popular platforms, including ASSISTments and Khan Academy [13], and there is mixed evidence as to whether BKT outperforms N-CCR at modeling student learning [8, 10, 13, 21]. Nevertheless, the scientific community shows a clear preference for BKT (and other more complex knowledge tracing algorithms) based on the allocation of research attention.

BKT is a two-state Hidden Markov Model where the unobserved hidden state being modeled is student learning, and for a given knowledge component, a student has a state of either learned or not learned [6, 17, 11]. Although BKT is already more sophisticated than N-CCR, critics have suggested that BKT is too simple of an algorithm for modeling human learning. They point to deep (neural network) learning models to better represent all factors that go into student learning [17, 11]. Mao and colleagues [17] found that deep learning models outperformed BKT on some learning tasks. However, they also acknowledge that these gains in performance might not be worth the loss in model interpretability. While researchers tend to consider BKT as one of the simpler and more explainable algorithms for knowledge tracing, practitioners and learners who are the end-users may not share this view.

The explainability of an algorithm, which is partly determined by how transparent, understandable, interpretable it is, can play an essential role in its adoption into applications. Barredo Arrieta and colleagues [1] identified these and other reasons for making algorithms more explainable: most relevant to the work on BKT are trustworthiness, confidence, causality, and accessibility. Prior research on algorithms in education has echoed this finding. Kizilcec [14] found that increasing transparency by providing users with additional information about an algorithm made users trust the algorithm more (though too much information can erode trust). Other studies have more specifically examined the interpretability of BKT in learning applications. Yeung [24] explored the use of Item Response Theory to make BKT and deep learning models more explainable, but they have not examined how users react to it. Zhou and colleagues [25] examined BKT explainability by creating visualization "explainables." They then designed an experiment to determine the effectiveness between a static and interactive visualization and found that the static explainable led to a better understanding of the BKT algorithm. More generally, research on Open Learning Models (OLMs) has advanced an understanding of how to visualize and explain learning models [3, 2]. OLMs provide users with interactive visualizations that grant them insights into learning algorithms, along with the ability to adjust the algorithm. This study will add to OLM research by expanding knowledge on how to explain and visualize information to foster positive attitudes.

The current study provides a foundational understanding of how individuals perceive BKT compared to N-CCR along several attitudinal dimensions, and how much verbal and visual explanations of BKT can improve those perceptions. Our review of prior work informed the following two hypotheses:

H1. Verbal and visual explanations of BKT lead participants to prefer it over N-CCR.

H2. Verbal and visual explanations of BKT will positively increase participants attitudes about the BKT algorithm.

3. METHODS

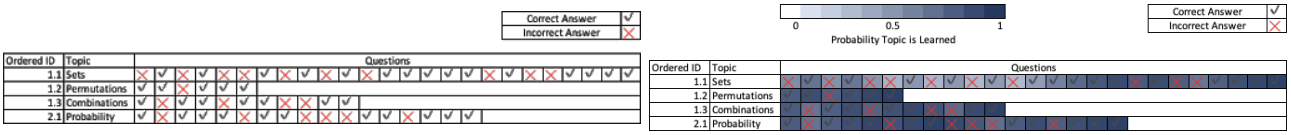
The study design, materials, measures and analysis approach are pre-registered with the Open Science Foundation: https://osf.io/7c5zt/. To refine the study design, measures, and analysis plan, we ran a pilot study with 26 participants and used both descriptive and inferential statistical analyses to build our analysis plan. We first used descriptive analysis to estimate survey completion time, ensure we had enough variance in responses, and check that the information provided to participants was enough information for them to evaluate the algorithms. We used respondents' answers and an open-ended question at the end of the survey in which we asked participants for any feedback to improve the survey. We took the results from this pilot study to alter the visualizations and information provided to participants and rephrase some questions to improve clarity. We removed the open-ended feedback question from the survey after the pilot.

3.1 Participants

Participants were recruited from Amazon Mechanical Turk and received $1.70 for completing a 10-minute survey. The study was advertised as seeking input on test preparation applications. To determine our target sample size of 170, we used G*Power to conduct a power analysis. Our analysis goals were to obtain 95% power to detect a medium effect size of 0.25 at the standard 0.05 alpha error rate with six repeated measures and four groups. While we had 170 participants who took the survey, 34 participants either failed to answer all of the comprehension questions correctly (29) or had prior experience with BKT (4) or both (1). Analyses were conducted on the remaining 136 respondents. Table 1 describes the sample demographics for the sample.

3.2 Procedure

To contextualize the study, participants were provided the following narrative with pictures of two sample questions taken from the ASSISTments platform:



(a) Simple visualization

(b) Detailed visualization

Figure 1: Two versions of a visualization of student performance on questions shown to participants depending on their condition assignment.

Table 1: Sociodemographics of Participants included in the study.

		n	%
Gender	Woman	79	58.1
	Man	54	39.7
	Transgender Man	1	.7
	Gender Variant/Non-Conforming	2	1.5
Ethnicity	Hispanic	10	7.4
	Not Hispanic	126	92.6
Race	White	100	73.5
	Black or African American	10	7.3
	American Indian or Alaska Native	2	1.5
	Asian	16	11.8
	Not Listed	5	3.7
	Multiracial	3	2.2
Age	18-24	27	19.9
	25-34	52	38.2
	35-44	37	27.2
	45-54	8	5.9
	55-64	11	8.1
	Above 65	1	.7

As an admissions requirement for a university program that you are applying for, you are preparing to take a general knowledge exam. The test is important to you and you need to do as well as possible to get accepted.
You have decided to use a test preparation app to help you study for the test.
A key feature of the test prep app is that it ***personalizes*** *the learning experience to help you study efficiently. The app shows you* ***only questions about topics that you have not already learned.***
The system keeps track of your answers to each question and ***automatically moves to the next topic once it determines that you have learned the previous topic.*** *To determine if you have learned a topic, the app uses an algorithm. Once the* ***algorithm determines you learned a topic****, it will stop giving you study questions about it. Thus, it also determines the speed at which you progress in your test prep.*
We would like to get your opinions about the ***two different algorithms*** *to understand which one you find more accurate and trustworthy.*

On the next page, participants answered three multiple-choice comprehension questions: (1) How does the test prep app determine what questions to give you? (2) What determines how quickly you are going to be done with test prep? (3) What happens when the system determines that you have learned a topic? We pre-tested these questions to ensure that an attentive reader would have no problems answering them correctly.

Next, participants saw a short description of the N-CCR algorithm, which we labeled as 3 Right in a Row (3RR): "A topic will be considered learned once a student correctly answers three questions in a row." A simple table depicting a sample student's progression for four topics (table rows) and questions for each topic (table columns) accompanied the description. The table looked like Figure 1a. Each cell contained an X or a ✓depending on if the student answered the question correctly. True to the 3RR algorithm, each topic was considered learned once three consecutive questions were answered correctly. At the bottom of the page, participants answered several questions about their attitudes towards the 3RR algorithm (see Measures).

At this point, participants were randomly assigned to conditions based on a 2x2 factorial design. There were 33 participants in the No BKT Explanation/BKT Simple Visualization condition in the final sample, 34 in the No BKT Explanation/BKT Detailed Visualization condition, 38 in the BKT Explanation/BKT Simple Visualization condition, and 31 in the BKT Explanation/BKT Detailed Visualization condition.

The following page mirrored the structure of the previous one but for BKT, providing a description and sample learning progress visualization based on the experimental assignment, followed by the same set of attitudinal questions about the algorithm. Next, on the final page of the survey, participants were asked to compare the two algorithms.

3.3 Experimental Manipulations

In the no BKT explanation condition, participants received this one-sentence description of the BKT algorithm: "A topic will be considered learned once the algorithm estimates with a high probability that a student has learned the topic based on their responses up to that point." In the BKT explanation condition, participants additionally received the following information about the BKT algorithm:

After every question you answer, the Bayesian Knowledge Tracing algorithm estimates the probability that you have now learned a topic using a probabilistic model that accounts for the following data:
– an initial probability that you have learned the topic based on your first answer: it is higher if you answered correctly
– a correct guess probability: e.g., 50% for a true/false question
– a slip probability for answering incorrectly even though you already learned the topic

– the difficulty of questions you have answered based on how many people have answered them incorrectly
– performance data such as the number of hints that you asked for and the time it took you to answer the question

Using all of this information, the algorithm estimates the probability that you have learned a topic. If the probability is above 95%, the algorithm moves you on to the next topic.

In the BKT simple visualization condition, participants received a simple table depicting a sample student's learning progress mirroring the one shown in the 3RR algorithm (Figure 1a). In the BKT detailed visualization condition, the same table was enhanced to show the estimated probability of having learned the topic using a color scale (Figure 1b). To make the visualizations realistic, we ran a BKT algorithm over a sample of ASSISTments data and used a 95% probability to determine mastery.

3.4 Measures

We measured participants' attitudes towards each algorithm using six items rated on 5-point unipolar response scales ('Not at all', 'Somewhat', 'Moderately', 'Very', 'Extremely'):

Confidence: "How confident are you that the test prep app with this algorithm will prepare you to do very well on the test?"
Understanding: "How well do you understand how this algorithm determines if you have learned a topic?"
Sophistication: "How complex is this algorithm for determining if you have learned a topic?"
Accuracy: "How accurate is this algorithm at determining if you have learned a topic?"
Trust: "How much do you trust this algorithm to determine what you have learned?"
Speed: "How quickly do you learn the materials for the test using this algorithm?"

At the end of the survey, participants rated their general preference over the two algorithms in response to the following question: "Now that you have learned about the 3 Right in a Row (3RR) and Bayesian Knowledge Tracing (BKT) algorithms, which one would you prefer to use for your test prep?" Response options were on a 7-point bipolar scale: 'Strongly prefer 3RR', 'Moderately prefer 3RR', 'Slightly prefer 3RR', 'Neither prefer 3RR nor BKT', 'Slightly prefer BKT', 'Moderately prefer BKT', 'Strongly prefer BKT'. Participants were invited to provide a rationale for their preference using an open-ended question: "Please tell us why you prefer the algorithm that you choose above."

3.5 Analytical Approach

We used the pilot study data to finalize our analysis plan by developing our inferential analysis. For H1, we decided to use linear regression to understand if the conditions had an association effect on the participants' overall preference. We used the conditions as the predictor variables and the preference as the outcome variable. We next decided to use multiple linear regression to understand the association between the attitudinal constructs and algorithm preferences. This analysis used the attitudinal constructs as the predictor variables with preference as the outcome variable. The last planned analysis evaluated H2 by running a linear regression on each attitudinal construct with the conditions as the predictor variables and the attitudinal construct as the dependent variable. While the interpretation of linear regression output is clear and familiar, we acknowledge that our measures are ordinal and not strictly continuous. We confirmed that analysis by ordinal logistic regression yields equivalent results.

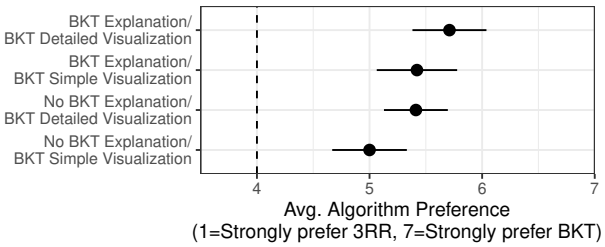


Figure 2: Average algorithm preference by condition.

For the open-ended question asking participants why they choose their preferred algorithm, we planned to use simple thematic coding. While we used the pilot data to create our analysis plan, we did remove all pilot data from the final dataset.

4. FINDINGS

First, we examine which algorithm participants preferred overall. Figure 2 shows their average preference in each condition, which varied between 5 (i.e. Slightly prefer BKT) and 6 (i.e. Moderately prefer BKT). While there is a suggestive pattern that providing more explanation for BKT strengthens the preference for BKT, this pattern was not statistically significant (linear regression: $F_{3,132} = 0.7455, p = 0.5268$). This means the data do not support H1.

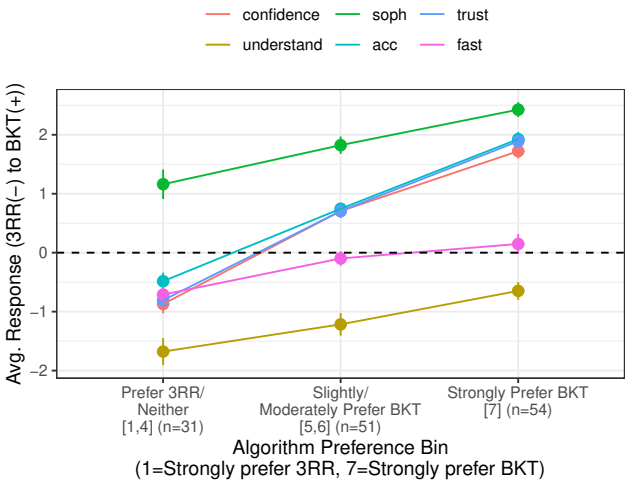


Figure 3: Average response on each measure at three levels of preference: Prefer 3RR to Neither, Slightly and Moderately Prefer BKT, and Strongly Prefer BKT. We choose these groupings because each group represents approximately 1/3 of the sample.

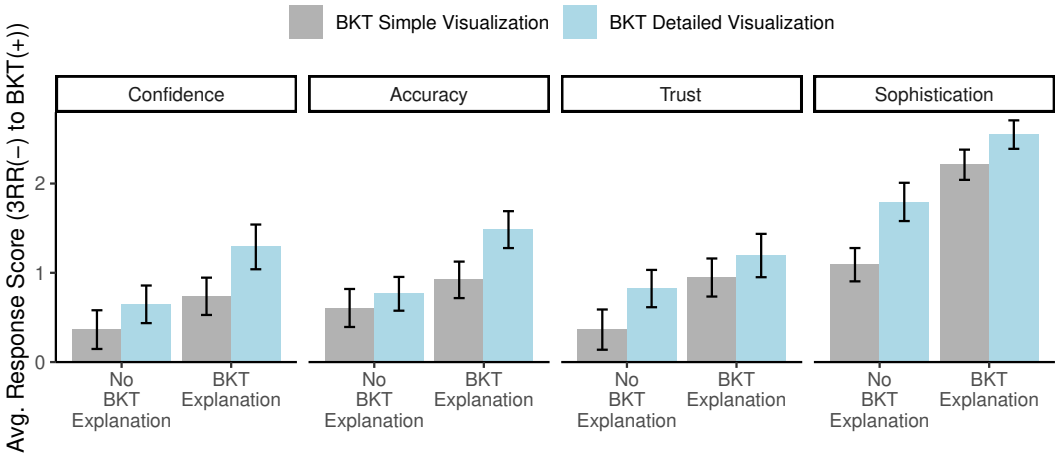


Figure 4: Average differences (BKT score - 3RR score) in significant attitudinal constructs as a function of the randomly assigned conditions. Positive scores indicate a higher score for BKT.

Next, we examine how algorithm preference is related to the six attitudinal measures: confidence in the learning system, the sophistication of the algorithm, trust, understanding, accuracy, and speed. We use the repeated measures design of our study by computing the difference score for each question: subtracting the participant's 3RR response from the BKT response. Figure 3 shows the average response on each measure at three levels of preference: Prefer 3RR to Neither, Slightly and Moderately Prefer BKT, and Strongly Prefer BKT. We choose these grouping because each group represents approximately 1/3 of the sample. All measures are positively correlated with preference as evidenced by their positive slopes (all Pearson's $r > 0.325, p < 0.0001$), but accuracy, confidence, and trust are correlated more strongly ($r > 0.739$). This highlights the importance of these three constructs in determining people's preference over the algorithms. In fact, the six measures explain 66.7% of the variance in preferences (multiple linear regression: $F_{6,129} = 43.07, p < 0.0001$).

Lastly, we examine how the provision of verbal and/or visual explanations influenced participant attitudes about BKT. Figure 4 shows the average response in each condition for the four measures that were significantly affected by the intervention (i.e., relative understanding and speed did not change significantly at $p < 0.1$). We find that confidence in the learning application with BKT (relative to 3RR) improved when both a detailed explanation and visualization were provided ($F_{3,132} = 2.88, p = 0.03844$). Likewise, the perceived accuracy of BKT improved with both types of explanation provided ($F_{3,132} = 3.28, p = 0.02305$). Trust in BKT improved by providing a detailed explanation, especially when complemented with the detailed visualization ($F_{3,132} = 2.346, p = 0.07575$). Finally, and not surprisingly, the more detail was provided, the more sophisticated BKT was perceived to be ($F_{3,132} = 11.17, p < 0.001$). This provides evidence in support of H2.

In all of our analyses, we tested for the presence of demographic heterogeneity in results. However, no significant demographic sources of variation were found in our sample.

5. DISCUSSIONS

This study investigated people's attitudes towards BKT relative to a more straightforward knowledge tracing algorithm and tested the effect of additional information via explanations and visualizations on their attitudes. Understanding how students might perceive the algorithms used in their learning applications is a crucial issue for the adoption and usability of these tools [9]. The results provide evidence supporting our second hypothesis that additional explanations improve key attitudinal measures of confidence, perceived accuracy, trust, and sophistication. Qualitative data from participants echo this result:

> For something high stake, I'd only trust the methods that employs a variety of learning modalities. The analytics for such should match the complexities of my learning process as well as the nature of the material I'm learning. The BKT would put me more at ease than the quick route of the 3RR approach. (Participant assigned to BKT Explanation and BKT Detailed Visualization who had high confidence, sophistication, accuracy, and trust in BKT relative to 3RR)

Surprisingly, we did not find a significant increase in people's preference for BKT (H1), even though we found that algorithm preference is explained largely by people's perceptions of accuracy, trust, and confidence. This preference for BKT regardless of experimental condition is furthered explained by the qualitative responses from participants:

> I don't believe the 3RR algorithm is at all beneficial to the student attempting to learn the topic. If the student just happens to get 3 exceptionally easy questions in a row, the algorithm will assume that the student has learned the topic which is likely not entirely true. (Participant assigned to No BKT Explanation and BKT Simple Visualization who Strongly Preferred BKT)

Nevertheless, participants generally preferred to use the BKT algorithm regardless of the experimental treatment.

Since confidence, trust, and accuracy are important to a user's preference for BKT, it was notable that those three measures were affected by the experimental manipulations. Consistent with prior studies of explainability and transparency in algorithmic systems, we also found that when more information about an algorithm is presented to people, they believe the algorithms to be more trustworthy and accurate, leading the user to have more confidence in the algorithm [1, 14]. This confidence can, in turn, increase the use of applications shown to improve educational outcomes. Our results further highlight the importance of OLM to provide more transparency to gain user trust and confidence. In addition, future educational applications using complex knowledge tracing algorithms should include detailed verbal and visualization explanations of the algorithm to improve confidence and trust in the application.

Frankly, we did not expect to find such strong support for BKT going into this study. Many learning platforms that have the capacity to implement BKT or even more complex algorithms have opted not to do so. Our informal understanding was that this was largely due to concerns that users, such as math teachers who use ASSISTments to assign homework, would not understand BKT and not trust it and lose confidence in the platform. This understanding led us to expect that participants would report a preference for the simpler algorithm and report that they find it more trustworthy and have more confidence in a system that uses it. Therefore, we are surprised by our positive findings for BKT and propose future research directions to follow up on these results.

Future work in this area should explore different participant populations and scenarios. We are planning to run this study with student samples to see if we can replicate the results. While we asked our participants to put themselves in the shoes of a student needing to prepare for a high-stakes test, running this experiment on a student population might make the scenario more realistic to the participants. Additionally, given that many tasks on Mechanical Turk involve subjects annotating machine learning datasets, this group of participants may have more favorable attitudes to algorithms. Another area for future work includes changing the scenario. We ran the study from the viewpoint of a student. However, we acknowledge that intelligent tutoring systems deployed in classrooms also need the trust and confidence of the teachers administering the applications. We plan to rewrite the scenario to allow participants to take on the role of a teacher.

6. REFERENCES

[1] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58(October 2019):82–115, jun 2020.

[2] S. Bull. There are open learner models about! *IEEE Transactions on Learning Technologies*, 13(2):425–448, 2020.

[3] S. Bull and J. Kay. Open learner models. In *Advances in intelligent tutoring systems*, pages 301–322. Springer, 2010.

[4] A. Bunt and C. Conati. Probabilistic student modelling to improve exploratory behaviour. *User Modeling and User-Adapted Interaction*, 13(3):269–309, 2003.

[5] H. Cen, K. Koedinger, and B. Junker. Learning factors analysis–a general method for cognitive model evaluation and improvement. In *International Conference on Intelligent Tutoring Systems*, pages 164–175. Springer, 2006.

[6] B. Choffin, F. Popineau, Y. Bourda, and J.-J. Vie. Das3h: Modeling student learning and forgetting for optimally scheduling distributed practice of skills. In *International Conference on Educational Data Mining (EDM 2019)*, 2019.

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

[8] Y. B. David, A. Segal, and Y. K. Gal. Sequencing educational content in classrooms using Bayesian knowledge tracing. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*, pages 354–363, New York, New York, USA, 2016. ACM Press.

[9] F. D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340, 1989.

[10] S. Doroudi and E. Brunskill. Fairer but Not Fair Enough On the Equitability of Knowledge Tracing. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 335–339, New York, NY, USA, mar 2019. ACM.

[11] T. Gervet, K. Koedinger, J. Schneider, T. Mitchell, et al. When is deep learning the best approach to knowledge tracing? *Journal of Educational Data Mining*, 12(3):31–54, 2020.

[12] W. J. Hawkins, N. T. Heffernan, and R. S. J. D. Baker. Learning bayesian knowledge tracing parameters with a knowledge heuristic and empirical probabilities. In S. Trausan-Matu, K. E. Boyer, M. Crosby, and K. Panourgia, editors, *Intelligent Tutoring Systems*, pages 150–155, Cham, 2014. Springer International Publishing.

[13] K. Kelly, Y. Wang, T. Thompson, and N. Heffernan. Defining Mastery: Knowledge Tracing Versus N-Consecutive Correct Responses. In *8th International Conference on Educational Data Mining*, 2015.

[14] R. F. Kizilcec. How Much Information? In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2390–2395, New York, NY, USA, may 2016. ACM.

[15] S. Klingler, T. Käser, B. Solenthaler, and M. Gross. On the performance characteristics of latent-factor and knowledge tracing models. *International Educational Data Mining Society*, 2015.

[16] K. R. Koedinger and V. Aleven. Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, 19(3):239–264,

2007.

[17] Y. Mao. Deep learning vs. bayesian knowledge tracing: Student models for interventions. *Journal of educational data mining*, 10(2), 2018.

[18] E. Millán and J. L. Pérez-De-La-Cruz. A bayesian diagnostic algorithm for student modeling and its evaluation. *User Modeling and User-Adapted Interaction*, 12(2):281–330, 2002.

[19] P. I. Pavlik Jr, H. Cen, and K. R. Koedinger. Performance factors analysis–a new alternative to knowledge tracing. *Online Submission*, 2009.

[20] R. Pelánek. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction*, 27(3):313–350, 2017.

[21] R. Pelánek and J. Řihák. Experimental Analysis of Mastery Learning Criteria. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 156–163, New York, NY, USA, jul 2017. ACM.

[22] S. Ritter, M. Yudelson, S. E. Fancsali, and S. R. Berman. How mastery learning works at scale. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pages 71–79, 2016.

[23] S. Schiaffino, P. Garcia, and A. Amandi. eteacher: Providing personalized assistance to e-learning students. *Computers & Education*, 51(4):1744–1754, 2008.

[24] C.-K. Yeung. Deep-IRT: Make Deep Learning Based Knowledge Tracing Explainable Using Item Response Theory. *EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining*, pages 683–686, apr 2019.

[25] T. Zhou, H. Sheng, and I. Howley. Assessing post-hoc explainability of the bkt algorithm. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 407–413, New York, NY, USA, 2020. Association for Computing Machinery.