Designing for Student Understanding of Learning Analytics Algorithms

Catherine Yeh[1], Noah Cowit[1], and Iris Howley[1]

Williams College, Williamstown, MA USA {cy3,nqc1,ikh1}@williams.edu

Abstract. Students use learning analytics systems to make day-to-day learning decisions, but may not understand their potential flaws. This work delves into student understanding of an example learning analytics algorithm, Bayesian Knowledge Tracing (BKT), using Cognitive Task Analysis (CTA) to identify knowledge components (KCs) comprising expert student understanding. We built an interactive explanation to target these KCs and performed a controlled experiment examining how varying the transparency of limitations of BKT impacts understanding and trust. Our results show that, counterintuitively, providing some information on the algorithm's limitations is not always better than providing no information. The success of the methods from our BKT study suggests avenues for the use of CTA in systematically building evidence-based explanations to increase end user understanding of other complex AI algorithms in learning analytics as well as other domains.

Keywords: Explainable artificial intelligence (XAI) · Learning analytics · Bayesian Knowledge Tracing.

1 Introduction

Artificial Intelligence (AI) enhanced learning systems are increasingly relied upon in the classroom. Due to the opacity of most AI algorithms, whether from protecting commercial interests or from complexity inaccessible to the public, there is a growing number of decisions made by students and teachers with such systems who are not aware of the algorithms' potential biases and flaws. Existing learning science research suggests that a lack of understanding of learning analytics algorithms may lead to lowered trust, and perhaps, lower use of complex learning analytics algorithms [26]. However, in different educational contexts, research has evidenced that increased transparency in grading can lead to student dissatisfaction and distrust [15], so more information is not guaranteed to be better. Furthermore, opening and explaining algorithms introduces different issues, as cognitive overwhelm can lead to over-relying on the algorithm while failing to think critically about the flawed input data [24]. As a first step toward realizing this relationship between algorithm, user understanding, and outcomes, this work answers the following questions about BKT:

- What are the knowledge components of algorithmic understanding for BKT?
- What factors impact successful learning with our interactive explanation?

– How does our explanation impact user attitudes toward BKT and AI?

To answer these questions, we first systematically identify the knowledge components (KCs) of BKT, then design assessments for those KCs. Next, we implement a post-hoc, interactive explanation of BKT using these KCs, evaluating our explanation in light of the assessments and user perspectives of the system. Finally, we run an additional experiment varying the amount of information participants are shown about BKT and measuring how this reduction in transparency impacts understanding and perceptions.

2 Prior Work

A student may use a learning analytics system displaying which skills they have mastered, and which they have not. This information can assist the student in determining what content to review, learning activities to pursue, and questions to ask. This information can also inform the learning analytics system of which practice problems to select, creating a personalized, intelligent tutoring system. Underneath the system display, there may be a Bayesian Knowledge Tracing algorithm predicting mastery based on student responses in addition to various parameters such as the likelihood of learning or guessing [2]. As an AI algorithm, BKT is inherently prone to biases and flaws [9]. Without a proper mental model for BKT, students may make decisions based on their own observations of system outputs, which may not be accurate. A sufficient understanding of the underlying algorithm may influence trust in the system as well as decision-making [14].

Researchers studying fairness, accountability, and transparency of machine learning (ML) view this topic from two angles: 1) where the ML algorithm can automatically produce explanations of its internal working, and 2) post-hoc explanations constructed after the ML model is built [19]. To achieve this second goal, researchers create **post-hoc explanations** to teach the concepts of particular algorithms [19]. Often, these explanations require the reader to have extensive prior knowledge of machine learning, despite a large proportion of algorithmic decision systems being used by non-AI/ML experts, such as healthcare workers and criminal justice officials. Additional ML work suggests using the basic units of cognitive chunks to measure algorithmic understanding [11]. A means to identify what knowledge experts rely on to understand complex algorithms is necessary to bridge this gap between post-hoc explanations for ML researchers and typical users, such as students with learning analytics systems.

Previous work has involved measuring and evaluating what it means to know a concept from within the learning sciences, but this question manifests itself somewhat differently in explainable AI (XAI) research. One approach is to adopt definitions from the philosophy of science and psychology to develop a generalizable framework for assessing the "goodness" of ML explanations. For example, in decision-making research, the impact of different factors on people's understanding, usage intent, and trust of AI systems is assessed via hypothetical scenario decisions [21]. Using pre-/post-tests to measure learning about AI algorithms is one possibility for bridging the ML and learning science approaches [27].

While increasing research explores different ways of evaluating post-hoc explanations in the ML community, it is not clear how XAI designers identify the concepts to explain. **Cognitive Task Analysis (CTA)** provides a rigorous conceptual map of what should be taught to users of ML systems by identifying the important components of algorithms according to existing expert knowledge [7].

Intelligent tutoring system design uses CTA to decompose content into the knowledge and sub-skills that must be learned as part of a curriculum [20]. Lovett breaks CTA down into 2x2 dimensions, the theoretical/empirical and the prescriptive/descriptive. Our study focuses on the *empirical/prescriptive* dimension of CTA, where a think aloud protocol is used as experts solve problems pertaining to the domain of interest (e.g., a particular algorithm). We chose to leverage a form of expert CTA, as studying expertise elucidates what the results of "successful learning" look like and what kinds of thinking patterns are most effective and meaningful for problem-solving [20]. Ultimately, these results from CTA can be used to design more effective forms of instruction for novices, such as explanations of learning analytics algorithms.

The knowledge and skills revealed by CTA are called **knowledge components** or **KCs**. KCs are defined as "an acquired unit of cognitive function or structure that can be inferred from performance on a set of related tasks" [16]. In this paper, we use CTA to systematically identify the different knowledge components that comprise the AI algorithm, Bayesian Knowledge Tracing, which are ultimately evaluated through observable assessment events.

Bayesian Knowledge Tracing (BKT) models students' knowledge as a latent variable and appears in Technology Enhanced Learning systems such as the Open Analytics Research Service [5]. BKT predicts whether a student has mastered a skill or not (either due to lack of data or low performance) using four parameters: `P(init)`, `P(transit)`, `P(guess)`, and `P(slip)`. In practice, these parameters are fit through a variety of methods [2] and may be shared across an entire class of students [9]. Additionally, `P(transit)`, `P(guess)`, and `P(slip)` are often not updated, remaining at their preset initial values [2]. BKT updates its estimates of mastery, `P(init)`, as a student proceeds through a lesson [2].

As a probabilistic algorithm, BKT falls subject to certain biases and limitations. For example, model degeneracy occurs when BKT does not work as expected due to its initial parameter values being outside an acceptable range [9]. BKT's parameters also do not account for certain events, such as forgetting [8] or the time it takes a student to answer a question, which are relevant and important to consider when assessing learning and mastery.

BKT is a sufficiently complex algorithm as to not be easily understood, but also sufficiently explainable as the parameters and how they interact are all known. While we use BKT as our algorithm of interest for this study, it is possible to apply these same methods of examination to other learning analytics algorithms that students and teachers may find difficult to understand.

3 Knowledge Components of BKT

We conducted a CTA to gain knowledge about expert understanding with respect to our selected AI algorithm, BKT. We examine BKT from the student perspective, as they represent one of BKT's target user groups. Our CTA protocol involves interviewing student experts of BKT and having them think aloud and step-through various scenarios that may be encountered when using a BKT system. This is analogous to the approach described in [20], which uses CTA for the design of intelligent tutoring systems for mathematics.

The participants in this study were seven undergraduate students at a rural private college who previously studied BKT as part of past research experiences and in some cases had implemented small-scale BKT systems. Interviews were semi-structured with a focus on responding to ten problems and lasted 30-60 minutes in duration. By recording comprehensive, qualitative information about user performance during these interviews [7], we were able to identify the knowledge components of student BKT expertise.

We developed our own BKT scenarios, as identifying problems for experts to solve is less straightforward than identifying problems for statistics experts to solve as in [20]. We adapted our approach from Vignette Survey design [1] to generate BKT problems within the context that experts ordinarily encounter. Vignette Surveys use short scenario descriptions to obtain individual feedback [1]. Additionally, the numerous social indicators of BKT parameters and weighing of subjective factors necessary for model evaluation make vignettes an optimal tool for this study. For instance, a lack of studying, sleep, or prior knowledge can all lead to a low starting value of `P(init)`.

For each scenario, we constructed a vignette describing background information about a hypothetical student followed by one or more questions regarding BKT (e.g., "Amari loves debating. They are very well spoken in high school debate club. Although Amari's vocabulary is impressive, they often have difficulty translating their knowledge into their grades. For example, Amari gets flustered in their high school vocab tests and often mixes up words they would get correct in debate. These tests are structured in a word bank model, with definitions of words given the user must match to a 10-question word bank. (1) What do you think are reasonable parameters for BKT at the beginning of one of these vocab tests? Please talk me through your reasoning..."). We were not only interested in comprehension of BKT's parameters and equations, but also the context in which BKT systems are used. Thus, our CTA protocol includes additional details such as test anxiety and other potential student differences that may create edge cases for interpretations of BKT output.

Our data was compiled after interviews were completed. First, the initial and final states (i.e., the given information and goal) for each scenario were identified. Questions with similar goals were grouped together, forming broader knowledge areas (e.g., "Identifying Priors"). Next, each participant's responses were coded to identify the steps taken to achieve the goal from the initial state. Then, we identified common steps used in each scenario. Final knowledge components were created by matching similar or identical processes from questions in the same

knowledge area. If a certain step was taken by the majority of participants but not all, we denote it as an "optional" KC by using *italics*.

We ultimately divided our analysis of BKT into four discrete but related knowledge areas: (1) **Identifying Priors**, (2) **Identifying Changed Parameters**, (3) **Evaluating P(init)**, and (4) **Limitations of BKT**. Each knowledge area consisted of 4-5 knowledge components, resulting in a total of 19 KCs:

Identifying Priors concerns the processing of subjective vignettes into reasonable numerical values for the four initial parameters of BKT.

1. Recall range of "normal values" and/or definitions for the parameter in question. This may involve recognizing (implicitly or explicitly) what `P(init / transit / guess / slip)` is and how it is calculated.
2. Synthesize (summarize or process) information from vignette, identifying specific evidence that is connected to the parameter in question.
3. *Consider BKT's limitations & how this could impact this parameter's value.*
4. Make an assessment about the parameter in question based on this qualitative evidence (or lack thereof).
5. Choose a parameter value by converting to a probability between 0 and 1.

Identifying Changed Parameters focuses on the direction of change in parameter values (if any).

1. Consider the prior parameter level of `P(init / transit / guess / slip)`.
2. Synthesize new information given, identifying specific evidence that suggests a change in parameter value (or a lack thereof).
3. Make an assessment about the parameter in question based on this qualitative evidence (or lack thereof).
4. Decide direction of change (increase, decrease, or stays the same).
5. If prompted, choose a new parameter value by converting assessment to a probability between 0 and 1.

Evaluating P(init) addresses how the parameter `P(init)` is essential for evaluating practical applications of BKT. Sometimes experts arrived at different answers for these more open-ended problems, but our participants typically followed similar paths to arrive at their respective conclusions.

1. Synthesize information from vignette, considering parameter level of `P(init)`.
2. Make a judgment as to the magnitude of `P(init)` (e.g., low, moderate, high, moderately high, etc.).
3. Consider magnitude with respect to the situation and BKT's definition of mastery. Some situations call for a very high level of knowledge—and thus a very high `P(init)` (e.g., space travel), while in other situations, a moderate level of knowledge is acceptable (e.g., a high school course).
4. Take a stance on the question. Often: "With this value of `P(init)`, has X achieved mastery?" or "...is 0.4 a reasonable value for `P(init)`?"
5. *Explain why BKT's predictions might not be accurate in this case due to its limitations, probabilistic nature, etc.*

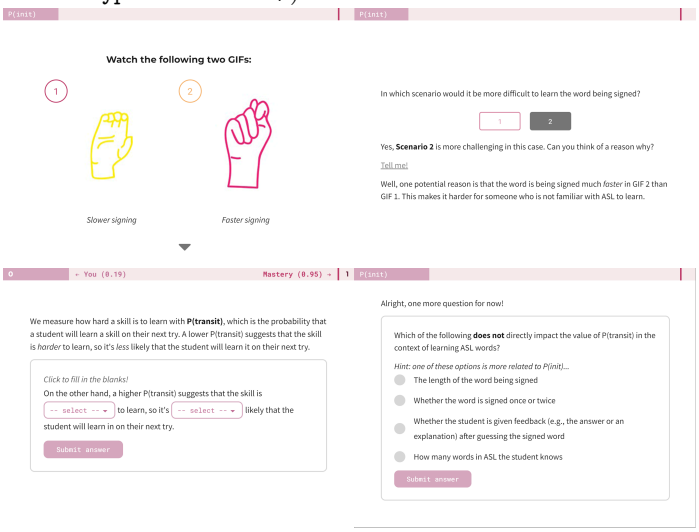
Limitations of BKT covers three limitations of BKT within this protocol: model degeneracy [9], additional non-BKT parameters such as time taken and forgetfulness between tests, and the probabilistic nature of BKT. In many cases, these problems also related to the "Evaluating P(Init)" knowledge area.

1. Synthesize information from vignette, identifying any "irregular" pieces of information (e.g., anything that's relevant to learning/mastery but not encompassed by the standard 4 BKT parameters, like whether a student is being tested before or after their summer vacation).
2. *If relevant, consider previous parameter values.*
3. Experiment with irregular information and consider limitations of BKT. This often involved asking open ended questions about learning/mastery.
4. Make a statement about BKT's analysis (correct or not correct, sensible/intuitive or not, etc.), or answer the posed question(s) accordingly, after determining that BKT does not account for this irregular information.

4 The BKT Interactive Explanation

With KCs established, we designed our BKT explanation using principles from user-centered design [1]. Following this iterative design process, we went through several cycles of brainstorming, prototyping, testing, and revising. Our final explanation is an interactive web application that uses American Sign Language to motivate and illustrate the behavior of BKT systems.

Fig. 1. P(transit) Module from BKT Explanation, artwork by Darlene Albert (`https://darlenealbert.myportfolio.com/`)



[1] `https://dschool.stanford.edu/resources/design-thinking-bootleg`

Along with the pedagogical principles of Backward Design [25], we made additional design decisions following best practices in learning and instruction, active learning, and self-explanation in particular. "Learning by doing" is more effective than passively reading or watching videos [17], and so our explanation is interactive with immediate feedback which research shows leads to increased learning. To ensure that we targeted the BKT KCs identified previously, we mapped each activity in our explanation to its corresponding KCs.

After learning about all four parameters, we bring the concepts together for a culminating mini game in which participants practice their ASL skills by identifying different finger-spelled words until they achieve mastery. Following the BKT mini game are four modules to teach BKT's flaws and limitations: (1) "When do you lose mastery?", (2) "What Causes Unexpected Model Behavior?", (3) "How do Incorrect Answers Impact Mastery?", and (4) "What is the Role of Speed in BKT?". This is essential to our goal of encouraging deeper exploration of the algorithm and helping users develop realistic trust in BKT systems. For example, in our "When Do You Lose Mastery?" module, participants are asked to assess the magnitude of `P(init)` at different points in time. This module demonstrates how BKT does not account for forgetting, which may bias its estimates of mastery.

To implement our interactive explanation, we created a web application coded in JavaScript, HTML, and CSS. We iteratively tested and revised our implementation with participants, until we reached a point of diminishing returns in which no new major functionality issues arose. The final design is a dynamic, interactive, publicly accessible explanation: `https://catherinesyeh.github.io/bkt-asl/`.

After the implementation phase, we assessed the effectiveness of our BKT explanation with a formal user study. We designed pre- and post-tests to accompany our BKT explanation in a remote format. Our pre-test consisted mainly of questions capturing participant demographics and math/computer science (CS) background. Prior work suggests that education level impacts how users learn from post-hoc explanations [27], so we included items to assess participant educational background and confidence. Math/CS experience questions were adapted from [13] using Bandura's guide for constructing self-efficacy scales [3]. We also collected self-reported familiarity with Bayesian statistics/BKT systems and general attitudes toward AI; these questions were based on prior work [10].

Our post-test questions were inspired by [22], which outlines evaluation methods for XAI systems. To evaluate user *mental models* of BKT, our post-test includes questions specifically targeting our BKT KCs, such as Likert questions like: "BKT provides accurate estimates of skill mastery." Many of our other questions were similar to the vignette-style problems included in our CTA protocol, mirroring the scenarios we present to participants throughout the explanation. To measure *usability and user satisfaction*, we adapted questions from the System Usability Scale [4] and similar scales [3,12,23]. We also measured *user trust* with a modified version of the 6-construct scale from [6]. Finally, we compared

user attitudes toward AI algorithms more generally before and after completing our explanation using the same set of questions from our pretest [10].

User study participants were nine undergraduate students from the same rural private college as above. Each participant completed a pre-test, stepped-through the explanation, and ended with a post-test. We do not go in-depth into this user study here, as our follow-up experiment uses the same measures with a larger sample size. Preliminary results suggest that any participant can learn from our BKT explanation regardless of their math/CS background, and that users received our BKT explanation positively. Satisfied with this initial evaluation, we moved to the next stage of this work: examining how the information in the interactive explanation impacts user understanding and other outcomes.

5 Impact of Algorithmic Transparency on User Understanding and Perceptions

With an effective post-hoc explanation of BKT, we can answer the following research questions: **RQ1:** How does decreasing the transparency of BKT's limitations affect algorithmic understanding? **RQ2:** How does decreasing the transparency of BKT's limitations affect perceptions of algorithmic fairness and trust? As these questions focus on trust and other user outcomes in decision-making situations, the most impactful factor of this is likely related to the user's understanding of the algorithm's limitations or flaws. And so, we designed a controlled experiment examining how three levels of information about BKT's limitations impact user perceptions of BKT vs. Humans as the decider of a hypothetical student's mastery in high-stakes and low-stakes evaluation circumstances.

To vary the amount of information provided about BKT's limitations, we included three explanation conditions. The *Long Limitations* condition included the original four limitations modules at the end of the BKT explanation. The *Short Limitations* condition reduced the multiple pages and interactive activities with a text summary and images or animations illustrating the same concepts. The *No Limitations* condition had none of the limitations modules. Participants were randomly assigned to an explanation condition.

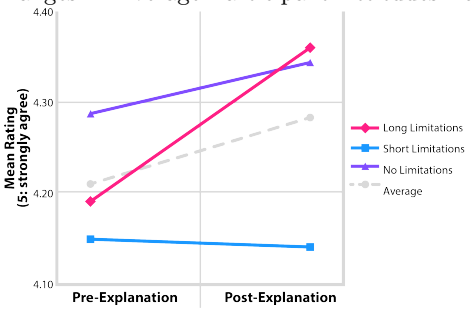
Similar to prior work on user perceptions of fairness of algorithmic decisions [18], we developed scenarios to examine algorithmic understanding's impact on user outcomes. In our case, we were interested in low-stakes and high-stakes situations involving BKT as the decision-maker, as compared to humans. Each scenario had a general context (i.e., "At a medical school, first-year applicants must complete an entrance exam. To be recommended for admission, applicants must score highly on this exam. An AI algorithm assesses their performance on the entrance exam.") and a specific instance (i.e., "Clay applies to the medical school. The AI algorithm evaluates their performance on the entrance exam."). These decision scenarios were added to our post-test previously described, along with measures from prior work asking about the fairness of each decision [18].

We recruited 197 undergraduate students from across the United States, 74 of which completed the pre- and post-tests satisfactorily. Of these, 50% identified

as female, 43% male, 5% other, and 2% did not respond. 82% reported being from the USA, with the remainder representing most other inhabited continents. 47% reported a major in math or engineering, and the rest were a mix of social & natural science, humanities, business, and communication. We later dropped 10 of these respondents due to outlying survey completion times or re-taking the survey after failing an attention check. 24 participants were assigned to the Long Limitations condition, 21 to Short, and 19 to No Limitations.

How does decreasing the transparency of BKT's limitations affect algorithmic understanding? There was no statistically significant relationship of time participants spent on the study and their post-test scores, nor was there a statistically significant difference of explanation conditions on time spent. Participants in the Long Limitations condition (μ=0.94, σ=0.09) had a higher average score on the **Limitations of BKT** knowledge area than the Short Limitations group (μ=0.92, σ=0.08), which in turn had a higher score than the No Limitations group (μ=0.88, σ=0.1). As understanding of BKT's limitations will depend on a more general understanding of BKT, we conducted a one-way ANCOVA to identify a statistically significant difference between explanation condition on learning in the Limitations of BKT knowledge area, controlling for performance in the three other knowledge areas, $F(3, 64) = 3.85$, $p < 0.05$. A Student's t-test shows that the No Limitations condition performed significantly worse on the Limitations of BKT knowledge area as compared to the other two explanation conditions. This suggests that our manipulation was mostly effective at impacting participant understanding of BKT's limitations.

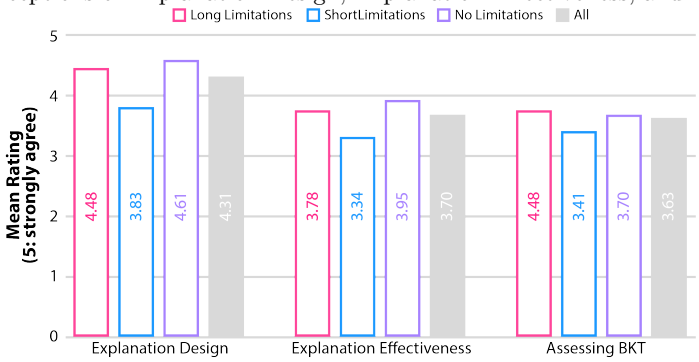
Fig. 2. Changes in Average Participant Attitudes Toward AI



How does decreasing the transparency of BKT's limitations affect perceptions of algorithmic fairness and trust? There was a significant interaction between experimental condition and initial attitudes about AI on final attitudes about AI, $F(2, 64) = 3.21$, $p < 0.05$, as shown in Figure 2.

All of our additional self-reported perceptions of the explanation design ($F(2, 64) = 10.77$, $p < 0.0001$), explanation effectiveness ($F(2, 64) = 4.89$, $p < 0.05$),

Fig. 3. Perceptions of Explanation Design, Explanation Effectiveness, and BKT Trust



and trust in the BKT algorithm ($F(2, 64) = 3.96$, $p < 0.05$) show a statistically significant effect of explanation condition. In all three of these cases, a Student's t-test shows that the Short Explanation condition has a significantly lower mean than the other two conditions, as shown in Figure 3.

We did not find significant results for the low/high-stakes X human/AI as the decision-maker questions. We likely need more than one question per category, or possibly longer exposure to BKT to measurably impact decision-making. However, in all cases, the Short Limitations condition had lower means than the other two conditions. This aligns with our prior results.

These results show that less information about an algorithm is not always worse. Participants in the Short Limitations group did not experience a positive increase in general attitudes about AI, unlike the Long and No Limitations groups, and the Short Limitations condition also perceived the BKT algorithm significantly less positively than the other two conditions. Despite the fact that the Short Limitations participants learned significantly more about BKT's limitations than the No Limitations condition, students in our middle-level information condition appear to have a significantly less positive perception of both our specific AI and AI more generally. For designers of interactive AI explanations, understanding how the design of an explanation impacts user perceptions is critical, and our work on student understanding of the learning algorithms they use provides a method for investigating that relationship.

6 Conclusion

This work shows that CTA can identify the necessary components of understanding a learning analytics algorithm and therefore, the necessary learning activities of an interactive explanation of the algorithm. Understanding the algorithm underlying learning analytics systems supports users in making informed decisions in light of the algorithm's limitations. Our CTA results identify four main knowledge areas to consider when explaining BKT: (1) **Identifying Priors**, (2) **Identifying Changed Parameters**, (3) **Evaluating P(init)**, and (4)

Limitations of BKT. We then varied the length of the limitations module in the implementation of an interactive BKT explanation. Results revealed that using a limitations module with reduced information can have surprising effects, mostly, a less than statistically expected impact on general perceptions of AI, as well as on perceptions of the learning algorithm itself.

Limitations of this work arise from limitations of the methods. The think aloud protocol for CTA shares limitations with all think aloud protocols: as a method for indirectly observing cognitive processes which are not directly observable, it is possible that some processes were missed by the think aloud protocol. Additionally, while these KCs apply to BKT, they may not generalize directly to another algorithm, although the CTA method itself certainly does extend to other contexts. The Short Limitations condition did not learn significantly less on the BKT limitations post-test as compared to the Long Limitations section, and so our results looking at explanation condition are likely capturing an effect based on more than just algorithmic understanding. Our posttest measures of high/low stakes human/AI decision makers only tested one decision scenario of each type, and needs to be expanded to be more generalizable. Furthermore, students are not the only users of learning analytics. Teachers are also important stakeholders, so next steps include repeating the process for instructor users of BKT. This information can be used to decide whether different explanations should be constructed for different stakeholders, or if a more general AI explanation could suffice, given that user goals are sufficiently aligned.

Our findings also inform future work involving other complex algorithms, with the larger goal of measuring how user understanding affects system trust and AI-aided decision-making processes. This process of applying CTA methods to identify important expert concepts that novices should learn about an algorithm, designing explanatory activities to target each KC, and then evaluating knowledge acquisition and shifts in decision-making patterns connected to each KC provides a generalizable framework for building evidence-based post-hoc AI explanations that are accessible even to non-AI/ML experts.

References

1. Atzmüller, C., Steiner, P.: Experimental vignette studies in survey research. Methodology **6**(3), 128–138 (2010)
2. d Baker, R.S., Corbett, A., Aleven, V.: More accurate student modeling through contextual estimation of slip & guess probabilities in bayesian knowledge tracing. In: Int Conf on Intelligent Tutoring Systems. pp. 406–415. Springer (2008)
3. Bandura, A.: Guide for constructing self-efficacy scales. In: Self-Efficacy Beliefs of Adolescents, pp. 307–337. Information Age, USA (2006)
4. Bangor, A., Kortum, P., Miller, J.: An empirical evaluation of the system usability scale. Int J of Human–Computer Interaction **24**(6), 574–594 (2008)
5. Bassen, J., Howley, I., Fast, E., Mitchell, J., Thille, C.: Oars: exploring instructor analytics for online learning. In: Proc of the ACM Conf on L@S. p. 55. ACM (2018)
6. Berkovsky, S., Taib, R., Conway, D.: How to recommend? user trust factors in movie recommender systems. In: Proc of the Int Conf on Intelligent User Interfaces. pp. 287–300. ACM (2017)

7. Clark, R., Feldon, D., van Merriënboer, J., Yates, K., Early, S.: Cognitive task analysis. In: Handbook of Research on Educational Communications & Technology, p. 577–593. Routledge, USA (2008)
8. Doroudi, S., Brunskill, E.: The misidentified identifiability problem of bayesian knowledge tracing. In: Proc of the Int Conf on Educational Data Mining. pp. 143–149. Int Educational Data Mining Society (2017)
9. Doroudi, S., Brunskill, E.: Fairer but not fair enough on the equitability of knowledge tracing. In: Proc of the Int Conf on LAK. pp. 335–339. ACM (2019)
10. Dos Santos, D.P., Giese, D., Brodehl, S., Chon, S., Staab, W., Kleinert, R., Maintz, D., Baeßler, B.: Medical students' attitude towards artificial intelligence: a multicentre survey. European radiology **29**(4), 1640–1646 (2019)
11. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017). https://doi.org/10.48550/ARXIV.1702.08608
12. Fu, F.L., Su, R.C., Yu, S.C.: Egameflow: A scale to measure learners' enjoyment of e-learning games. Computers & Education **52**(1), 101–112 (2009)
13. Hutchison, M., Follman, D., Sumpter, M., Bodner, G.: Factors influencing the self-efficacy beliefs of first-year engineering students. J of Engineering Education **95**(1), 39–47 (2006)
14. Khosravi, H., Shum, S.B., Chen, G., Conati, C., Tsai, Y.S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., Gašević, D.: Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence **3**, 100074 (2022)
15. Kizilcec, R.: How much information? effects of transparency on trust in an algorithmic interface. In: Proc of the SIGCHI Conf on Human Factors in Computing Systems. pp. 2390–2395. ACM (2016)
16. Koedinger, K., Corbett, A., Perfetti, C.: The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. Cognitive Science **36**(5), 757–798 (2012)
17. Koedinger, K., Kim, J., Jia, J.Z., McLaughlin, E., Bier, N.: Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In: Proc of the ACM Conf on Learning at Scale. pp. 111–120. ACM (2015)
18. Lee, M.K.: Understanding perception of algorithmic decisions: Fairness, trust, & emotion in response to algorithmic management. Big Data & Society **5**(1) (2018)
19. Lipton, Z.: The mythos of model interpretability. ACM Queue **16**(3), 31–57 (2018)
20. Lovett, M.: Cognitive task analysis in service of intelligent tutoring system design: A case study in statistics. In: Int Conf on ITS. pp. 234–243. Springer (1998)
21. Lu, J., Lee, D., Kim, T.W., Danks, D.: Good explanation for algorithmic transparency. In: Proc of the AAAI/ACM Conf on AIES. p. 93. ACM (2020)
22. Mohseni, S., Zarei, N., Ragan, E.: A multidisciplinary survey & framework for design & evaluation of explainable ai systems. ACM TiiS **11**(3-4), 1–45 (2021)
23. Phan, M., Keebler, J., Chaparro, B.: The development & validation of the game user experience satisfaction scale. Human Factors **58**(8), 1217–1247 (2016)
24. Poursabzi-Sangdeh, F., Goldstein, D., Hofman, J., Wortman Vaughan, J., Wallach, H.: Manipulating & measuring model interpretability. In: Proc of the SIGCHI Conf on Human Factors in Computing Systems. pp. 1–52. ACM (2021)
25. Wiggins, G., Wiggins, G., McTighe, J.: Understanding by Design. Association for Supervision & Curriculum Development, USA (2005)
26. Williamson, K., Kizilcec, R.: Effects of algorithmic transparency in bkt on trust & perceived accuracy. Int Educational Data Mining Society (2021)
27. Zhou, T., Sheng, H., Howley, I.: Assessing post-hoc explainability of the bkt algorithm. In: Proc of the AAAI/ACM Conf on AIES. pp. 407–413. ACM (2020)