Williams College

Yeh, Catherine

2022 Computer Science Thesis

Toward an Empirical Framework for Post-hoc Explainable AI

Advisor	Iris Howley
Additional Advisor	
Access	None of the above
Contains Copyrighted Material?	No
Release Restrictions	release now
Authenticated Access	

Toward an Empirical Framework for Post-hoc Explainable AI

by
Catherine Yeh

Professor Iris Howley, Advisor

A thesis submitted in partial fulfillment
of the requirements for the
Degree of Bachelor of Arts with Honors
in Computer Science

Williams College
Williamstown, Massachusetts

Contents

List of Figures

List of Tables

Abstract

Users of artificial intelligence (AI) decision-making systems rely on algorithms to help them make day-to-day decisions, but may not understand their potential flaws and biases due to algorithmic opacity. Many of these decisions have dire consequences, particularly in high-stakes fields such as criminal justice and healthcare. A popular method for increasing algorithmic transparency and subsequent user understanding,is creating post-hoc explainables, which use interactive visualizations to teach end users about AI techniques and algorithms. However, there is no clearly defined, systematic way for explainable designers to decide which concepts are necessary to teach, and many of their resulting explanations still require extensive prior knowledge of machine learning (ML) although users of these AI-mediated systems are typically not AI/ML experts.

Thus, we are interested in developing a framework for more robust, evidence-based explainable AI (XAI). My thesis builds off previous research that used a method from the learning sciences and human-computer interaction called Cognitive Task Analysis (CTA) to rigorously identify the necessary knowledge components (KCs) that comprise expert understanding of complex algorithms. To pilot this approach, we applied CTA to Bayesian Knowledge Tracing (BKT) – an AI algorithm commonly used in learning analytics systems to predict skill mastery – and identified four knowledge areas with 19 KCs.

Following the principles of Backward Design, I created pre- and post-tests to assess these KCs determined via CTA. Next, I designed and implemented a BKT explainable that targets each KC to evaluate the post-hoc interpretability of BKT and measure the impact of algorithmic understanding on behaviors of non-expert student users of BKT systems. This explainable uses American Sign Language to teach BKT, demonstrating the real-world applicability of BKT, and shows the algorithm in action on a smaller scale. Our explainable design process was heavily inspired by user-centered design and strives to follow best practices from teaching and learning theory. We also include additional modules about BKT's flaws and biases to encourage deeper exploration of the algorithm.

After the implementation phase, we ran user studies with our pre- and post-tests to evaluate the efficacy of my final BKT explainable. Our participants all demonstrated satisfactory performance on the post-test and exhibited sophisticated comprehension of BKT's limitations, providing evidence of the promise of our novel empirical XAI framework. The success of the methods from this work suggests potential avenues for the use of CTA, Backward Design, learning theory, and user-centered design in systematically building evidence-based, post-hoc explanations to increase end user understanding of other complex algorithms.

Acknowledgments

To my family and friends: a huge thank you for supporting and believing in me throughout this amazing journey and for always being willing to help pilot test my work. I am also very thankful to all the students at Williams and beyond who volunteered to participate in my user studies and the wonderful CS professors who I have been fortunate to get to know over the past four years. Thank you to my fellow HAI research assistants (especially Mira Sneirson '22 and Noah Cowit '20) for your contributions to this project as well—it was incredible to have the opportunity to work with you and this thesis would not be here without your help!

I would also like to thank my fantastic second reader, Rohit Bhattacharya, for his support and for always giving me the most insightful feedback and advice. Finally, I would like to express my endless gratitude and appreciation for my amazing thesis advisor, Iris Howley. I am so grateful to have gotten to work with Iris since my freshman summer and cannot thank her enough for all the mentorship, guidance, and encouragement she has offered me throughout my time at Williams. Working with Iris has truly inspired my love for research and I look forward to carrying her teachings with me wherever I go in life.

Chapter 1

Introduction

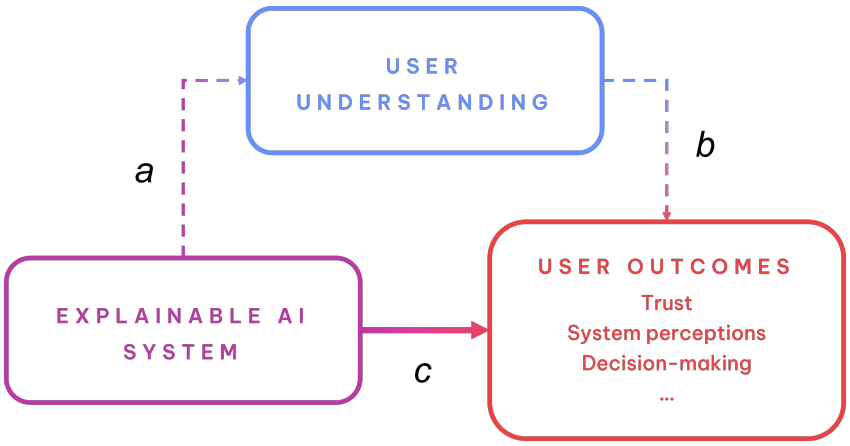
1.1 The Problem

Nowadays, more and more people are relying on artificial intelligence (AI)-enhanced systems for everyday decision-making. AI algorithms are being used to predict medical diagnoses [23], perform risk assessments in criminal justice systems [61], decide who gets approved for bank loans [9], and everything in between. However, many of these AI-mediated systems are inaccessible and obscured from the public due to algorithmic opacity, whether due to the proprietary nature of the algorithms to protect commercial interests, or due to the technical complexity of the algorithms themselves. Additionally, most users of such systems are not AI or machine learning (ML) experts. Together, these two facts mean that there is a growing number of decisions made by users of these ML systems who may not understand their potential flaws and biases. This can lead to dire consequences, particularly when AI decision-making systems are used in high-stakes fields such as healthcare and criminal justice [61].

Increasing algorithmic transparency has been proposed as one way to instill more realistic trust in AI systems [7] and allow for more reliable decisions to be made with the assistance of ML models [35]. This has led to increasing interest in the field of explainable AI (XAI), which involves utilizing various means of explanation to elucidate complex algorithms for the general public. With XAI, we can improve user understanding of AI systems, alleviating user confusion and frustration while instilling an appropriate level of system trust.

However, increasing algorithmic transparency alone is not guaranteed to generate positive user outcomes. Previous work has found that the presence of explanations does not necessarily improve the performance of human-AI decision-making teams [14] and misleading AI explanations can easily manipulate user trust [61]. To gain a clearer picture of how AI transparency influences user trust and other behaviors, we must examine what types of transparency impact user outcomes, and how. Different types of XAI may provide different perspectives of algorithmic understanding to users, which may mediate the relationship between the system and changes in user behavior. This idea is illustrated in Figure 1.1, where arrow c represents the direct path from XAI to behavioral outcomes, and arrows a and b demonstrate how user understanding could serve as a mediator.

Figure 1.1: Algorithmic Understanding as a Mediator of XAI Systems and User Outcomes



Another problem is that currently, XAI designers have varied methods of deciding which concepts are necessary to teach [69, 107], and most are not well-documented. There is also no established way to evaluate whether an explainable is "successful." Due to this lack of methodical approach, many AI explanations still require extensive prior knowledge of machine learning, making them inaccessible to most users. These findings motivate the need to better understand what constitutes a "good" explanation so we can design better and more systematic XAI to help instill more realistic trust in ML systems.

1.2 Goals & Plan

The goal of my thesis is to design a more robust, evidence-based XAI framework that addresses the issues discussed above. To do this, I apply existing methods from human-computer interaction, the learning sciences, and educational psychology to develop more effective explanations for AI algorithms. In particular, we focus on post-hoc AI explanations [67], which are constructed after model training. Our overall approach is based on Backward Design [109], which in the context of post-hoc XAI, involves first determining what the user should be able to do after completing the explanation, then creating metrics to determine whether those objectives are achieved, and ending with the design of the learning activity itself.

To begin this process, we used Cognitive Task Analysis (CTA), a method from human-computer interaction and the learning sciences for rigorously identifying necessary knowledge components (KCs) contributing to user understanding of an AI system. CTA can be used to study many algorithms and systems, but for the purposes of this project, we focus on applying CTA to Bayesian Knowledge Tracing (BKT), an AI algorithm used in learning analytics systems for predicting student mastery of skills [25]. Many educational technologies are built on top of algorithms like BKT, making students an important target user population for our research on algorithmic understanding. But as [61, 58] imply, even if a complete understanding of BKT were instilled in students, it is unclear how this would impact their interactions with and sentiment toward their educational technology.

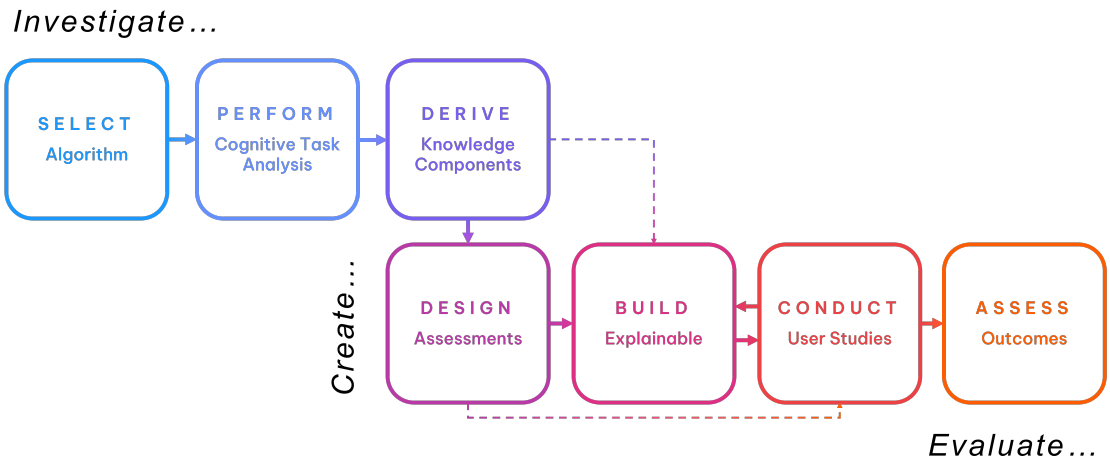


Figure 1.2: Our Proposed Framework for Designing Evidence-Based Post-Hoc AI Explainables

As a first step toward realizing the relationships in Figure 1.1, my work strives to answer the following questions:

- What are the knowledge components of algorithmic understanding for BKT?
- Can we build an explainable to teach these BKT concepts?
- What factors impact successful learning with our explainable?
- How does our explainable impact user attitudes toward BKT and AI more broadly?

We aim to answer these questions by applying a suite of empirical methods from design and learning theory: Backward Design, CTA, user-centered design, and best practices in teaching & learning. Prior to this thesis, we systematically identified the KCs of BKT using CTA. Next, I will design pre- and post-assessments for those KCs. Then, I plan to implement a post-hoc explanation for BKT to explain these concepts. Finally, I intend to evaluate my explanation with a user study, using the assessments previously designed, as shown in Figure 1.2.

Ultimately, we hope to champion a new, evidence-based explainable design process. If successful, this framework of determining the necessary KCs for understanding an algorithm via CTA methods and then designing post-hoc explanations to evaluate the impact of user understanding on decision-making processes can be generalized to other complex AI or ML algorithms as well.

1.3 Thesis Outline

The remainder of my thesis is organized into chapters as follows:

- **Chapter 2** provides background information and related work on the key topics discussed in my thesis, including *Explainable AI (XAI)*, which we are designing a framework for, *Cognitive*

Task Analysis (CTA), one of the key methods we use to create our XAI framework, and *Bayesian Knowledge Tracing (BKT)*, the algorithm we test this framework on.

- **Chapter 3** describes the prior work done by myself and other research assistants working with Prof. Howley leading up to this thesis. This chapter includes our CTA protocol we ran with student BKT experts to derive the necessary knowledge components to teach stakeholders of the algorithm, which serves as the basis for our resultant post-hoc explainable.

- **Chapter 4** discusses the various learning sciences and human-computer interaction inspired methodologies used throughout my thesis to create and evaluate our BKT explainable. This chapter is broken down into three sections: *explainable design*, *explainable implementation*, and *explainable evaluation*.

- **Chapter 5** details the results from our formal user studies in terms of evaluating the effectiveness of my final BKT explanation. I also talk about correlations in the data (e.g., time-on-task vs. learning) and how algorithmic understanding impacts user attitudes towards BKT and AI systems more broadly.

- **Chapter 6** presents an additional study completed to further investigate the impact of varying algorithmic understanding on user perceptions of fairness and trust. In this chapter, I share the motivation, methods, and preliminary results for this work, introducing three additional research questions:

  - Are the results from our first study generalizable to other college students?
  - How does decreasing the transparency of BKT's limitations affect algorithmic understanding?
  - How does decreasing the transparency of BKT's limitations affect perceptions of algorithmic fairness and trust?

- **Chapter 7** offers concluding remarks. Specifically, I discuss the *key contributions* of my thesis and outline some directions for *future work*, which may involve assessing the wider applicability of our XAI framework and developing behavioral measures to more directly observe the impact of algorithmic understanding on user outcomes.

Chapter 2

Background

In this chapter, I will provide an overview of related work and background information relevant to the goals and scope of my thesis. First, I will motivate the need for explainable AI (XAI) and outline the challenges in crafting robust, efficacious explanations for AI-mediated systems. Next, I will discuss how a learning sciences perspective may aid XAI designers and how methods such as Cognitive Task Analysis can help us systematically determine the necessary concepts needed to understand complex AI algorithms. Finally, I will introduce Bayesian Knowledge Tracing, our target algorithm in this work.

2.1 Explaining AI Algorithms

As the prevalence of AI-mediated decisions-making systems grows, so has the field of **explainable AI (XAI)**. XAI methods strive to make AI algorithms more transparent for the general public through various means of explanation, as most users of these systems have little to no experience with artificial intelligence/machine learning. AI explanations are important and valuable in many ways; they can inform stakeholders of use cases that are "out of scope," justify why specific performance measures were chosen to assess the model over others, list ethical considerations when using AI systems, and reveal other inherent flaws or biases [72].

Previous work has also found that increasing algorithmic transparency can lead to more realistic trust in ML models [7] and allow for more reliable decisions to be made with the assistance of these ML systems [35]. However, misleading AI explanations can easily manipulate user trust [61] and increasing algorithmic transparency is not guaranteed to generate positive user outcomes. For example, XAI does not necessarily improve the performance of human-AI decision-making teams [14] nor reduce over reliance on AI systems [22]. Within educational contexts, increased transparency in grading has also been shown to decrease student trust and satisfaction [58]. Completely opening and explaining algorithms also introduces difference issues, as cognitive overwhelm can lead to users over-relying on AI systems while failing to think critically about flaws in the input data [89]. To gain a clearer picture of how AI transparency influences trust and other behaviors, we must examine what and how different types of XAI affect user outcomes, as shown in Figure 1.1.

2.1.1 Explainability vs. Interpretability

Before diving deeper into our discussion of XAI, I wanted to briefly address the question of explainability vs. interpretability. While these terms are often used interchangeably by researchers, there are subtle differences between them. On one hand, **interpretability** is mostly concerned with elucidating the causal relationships within a ML system [74]—for example, which outputs are associated with which inputs. On the other hand, **explainability** focuses more specifically on understanding the internal processes of ML systems and *why* the model makes certain decisions or predicts certain outputs from its given inputs [66]. Thus, we generally consider fields such as interpretable machine learning [74, 95] to be broader than XAI, but it is also true that interpretable models are not necessarily explainable, and vice versa [35, 66].

My thesis focuses more around XAI as we are interested in providing users with a deeper, more nuanced understanding of model behavior beyond inputs and outputs. Explainability also is better suited to our goal of encouraging people to grapple with the limitations and flaws of ML algorithms and helping users develop appropriate levels of system trust. However, interpretability is still an important part of the discussion about transparentizing AI models, and this term will be included in relevant sections of my thesis. In many cases, establishing model interpretability is a natural and necessary first step to establishing model explainability [66].

2.1.2 Generating Explanations

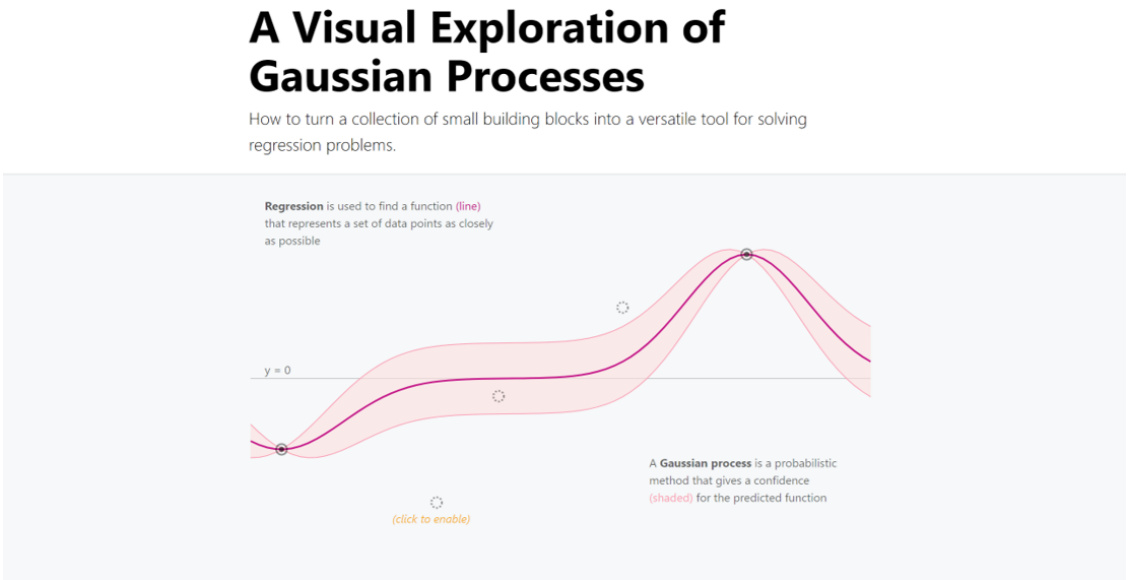
Since it seems that different types of XAI may have different impacts on user behaviors/outcomes, a natural first question to ask is, "What makes a good explanation?" Prior research has relied heavily on rationale-based methods to evaluate the "goodness" of an AI explanation, drawing from fields such as philosophy and psychology [69, 107]. But how do we go about transparentizing AI algorithms toward designing this "better" XAI? This is undoubtedly a challenging endeavor, as many machine learning algorithms are "black boxes" and not easily made transparent through current methods of XAI that strive to automatically produce explanations of an algorithm's internal works [107]. Even when an AI system is less opaque and more openly available to investigate, users without a background in AI/ML may still be unable to comprehend the underlying algorithm.

In these cases, one popular approach to increasing algorithmic transparency, and ultimately user understanding, is **post-hoc interpretability** [67, 95], which involves creating explanations for system predictions after model training. By contrast, intrinsic interpretability involves building "self-explanatory" models that are interpretable simply due to their inherent structures (e.g., short decision trees) [36]. My work focuses on the former notion of *post-hoc* interpretability.

For example, many researchers are beginning to teach the concepts of particular algorithms using post-hoc **explainables** "that explain how AI techniques work using visualizations" [67, 105]. Explainables often take the form of interactive graphs or visualizations interspersed with paragraphs of explanatory text, as shown by the examples in Figure 2.1 [105]. To expand these explainables and see them in action, click on the captions below each figure.

However, many of these explainables still require extensive prior knowledge of machine learning, despite most users being non-AI/ML experts, such as healthcare workers [23], loan officers [9], and

(a) Beginner's Guide to Dimensionality Reduction



(b) A Visual Exploration of Gaussian Processes

Figure 2.1: Example Explainables from the Visualization Community

criminal justice officials [61]. Different kinds of visualization techniques may also impact perceived fairness of ML predictors and algorithms [101]. Without a complete understanding of how an algorithm works, users may be vulnerable to the model's unquantified biases, particularly in problem contexts involving scientific understanding, safety, ethics, mismatched objectives, or multi-objective trade-offs [35]. Another challenge is that different stakeholders of AI systems have different needs and goals [73, 99]. Thus, explainables may have to target various forms of model interpretability.

2.1.3 A Learning Sciences Perspective on Explanation

One of the biggest problems with current interpretability research in XAI is that it largely misses the modern literature on teaching and learning. If we imagine the user as a learner and the post-hoc interpretation as the learning content, this would allow researchers to leverage the entire body of educational psychology and learning science research to achieve their goals of creating post-hoc explanations of complex algorithms. For example, in the visualization community, there has been a push to start viewing communicative visualizations as a learning problem [2] and taking advantage of techniques from education psychology such as inquiry-based learning [39].

In terms of evaluating model interpretability, recent work has looked at whether users understand ML definitions of fairness using measured performance on a series of questions [94]. However, this is a different, albeit related, task to assessing AI explanations. Here, participant understanding is measured through performance on a series of problems, which the researchers call a "comprehension score." Similarly, [114] uses pre- and post-test questions to measure learning comprehension scores about an AI algorithm. In [6], participant understanding is assessed through their ability to describe and predict reinforcement learning agent behavior in real-time strategy (RTS) games. Researchers have also used the RTS domain to evaluate different XAI methods, drawing from Information Foraging Theory [86] to better understand how humans navigate information seeking processes [84].

Other studies have investigated the impact of different factors on people's perceptions of understanding, usage intention, and trust in AI systems via hypothetical scenario decisions by participants [69]. This work proposes that the "goodness" of AI explanations is determined by how they meet epistemic, ethical, or consumer objectives. Researchers are also creating human-centered XAI frameworks that consider user goals for explanations, such as filtering, generalization, and impacting trust, as well as potential user cognitive biases [107]. Nonetheless, such frameworks are still too high-level to apply directly to individual AI explanation design scenarios [107]. My work takes a first step toward filling in the holes of these theory-based frameworks with empirically-based, learning science inspired methods that can be leveraged to aid in specific design decisions when constructing XAI.

Even as more research is being done to explore different ways of evaluating explainables in the ML community, it remains unclear how explainable designers identify the concepts necessary to teach stakeholders. A learning science perspective would provide a rigorous map of concepts that could be taught to users of ML systems, mirroring the approach taken in [39] for scientific visualization technologies. Portions of this conceptual map could then be evaluated against desirable outcomes, whether that be increased comprehension, shifts in decision-making, or even changes in user behavior regarding data investigation.

Table 2.1: Dimensions of Cognitive Task Analysis

	Theoretical	Empirical
Prescriptive	Information-theoretic analysis of how the task should be performed	Analysis of how experts produce high-quality solutions
Descriptive	Complexity analysis to predict how people actually do the task (e.g., typical errors)	Analysis of how novices actually do the task (e.g., their misconceptions)

2.2 What Does it Mean to Understand an Algorithm?

The first step in designing better explainable AI is determining what information should be provided —that is, what is important for stakeholders to know and what constitutes "understanding" for a particular algorithm. Historically, these types of questions have mainly been investigated in the field of computer science education, for example to assess how LOGO's turtle graphics impact students' systematic thinking skills and understanding of geometric concepts [92], and have not focused around studying specific algorithms. But even when algorithms are involved, most of this research considers and evaluates understanding from a higher-level course learning objective abstraction. For example, [71] use a game-based project to teach the A* algorithm, a fundamental AI search technique, toward their goal of providing students with a "significant learning experience." To truly grasp all the individual conceptual and procedural pieces of knowledge necessary to understand a specific algorithm, we must look to methods from the cognitive and learning sciences.

2.2.1 Cognitive Task Analysis

In this work, we use **Cognitive Task Analysis (CTA)**, an approach from the learning sciences and human-computer interaction commonly used to systematically analyze how individuals complete complex tasks through a combination of interview and observation methods [24, 108, 24]. CTA has been previously used to decompose learning curricula into the knowledge and sub-skills that should be taught to students in intelligent tutoring systems [68]. As seen in Table 2.1, CTA can be broken down into 2x2 dimensions: the *theoretical/empirical* and the *prescriptive/descriptive* [68].

The *theoretical/prescriptive dimension* involves a researcher identifying each step needed to move from an initial state to a final state that fulfills the requirements of a particular task. For example, if the task is figuring out how to conduct and interpret a statistical analysis to answer a question about a given data set, the first intermediary step would be identifying the types of variables involved to determine the appropriate type of analysis to perform [68]. Additional sub-steps could be identified through further CTA in this manner. In comparison, the *theoretical/descriptive* dimension considers how people will actually complete the task, not just "the correct way" to do it. Thus, these CTAs would also include common errors or misconceptions.

Our study focuses on the *empirical/prescriptive* dimension of CTA, where a think aloud protocol is employed as experts solve a series of problems pertaining to the domain of interest (e.g., a particular algorithm) [68]. The *empirical/descriptive* dimension is similar, but think alouds are instead conducted with novices to observe how tasks are accomplished by non-experts. We chose to leverage a form of expert CTA, as studying expertise can elucidate what the results of "successful learning" look like and what kinds of thinking patterns are most effective and meaningful for cognitive problem-solving [26]. Additionally, by comparing how people should reason/explain with how they actually do, including errors, XAI designers can support realistic reasoning processes while mitigating common cognitive biases [107]. Ultimately, these results from CTA can be used to design more effective forms of instruction for novices, such as AI explainables.

2.2.2 Assessing Understanding

The knowledge and skills ultimately revealed by Cognitive Task Analysis are called **knowledge components** or **KCs**. A KC is formally defined as "an acquired unit of cognitive function or structure that can be inferred from performance on a set of related tasks" [59]. However, as KCs themselves are not directly observable, developing from unobservable learning events, they should be evaluated through observable instructional and assessment events to environmentally validate the results of CTA, as outlined in the Knowledge-Instruction-Learning (KLI) framework [59]. One way to assess KCs is through a Backward Design approach [109], which is commonly used in the educational community. In the context of designing post-hoc AI explanations, Backward Design may involve developing assessments to evaluate whether system users have acquired the identified KCs, followed by the creation of explainables or transparency modifications to the algorithm of interest to target each identified KC.

With XAI, we consider the act of explaining to be teaching. Essentially, in Backward Design, the pedagogical (or explainable) designer starts with the end: what the user/student should be able to do by the end of learning experience or explanation, followed by creating metrics to determine whether those objectives were achieved, and ending with the design of the learning activity itself [109]. This strategic design process ensures that the resultant post-hoc explanation aligns with its pre-determined end goals from the perspective of user abilities, while also providing a means for evaluating success. After all, only after the appropriate instructional content is developed can we more definitively assess how algorithmic understanding impacts user decision-making and trust in AI-mediated systems, which are idealized outcomes in the Fairness, Accountability, and Transparency (FAccT) machine learning model [1].

2.3 Bayesian Knowledge Tracing

We chose **Bayesian Knowledge Tracing (BKT)** as our algorithm of interest for this research. As a probabilistic algorithm, BKT inherently comes with certain limitations and flaws, making it sufficiently complex as to not be easily understood. On the other hand, because its parameters and how they interact are all already known, BKT is simultaneously a relatively approachable and

Figure 2.2: Bayesian Knowledge Tracing as a 2-state Hidden Markov Model

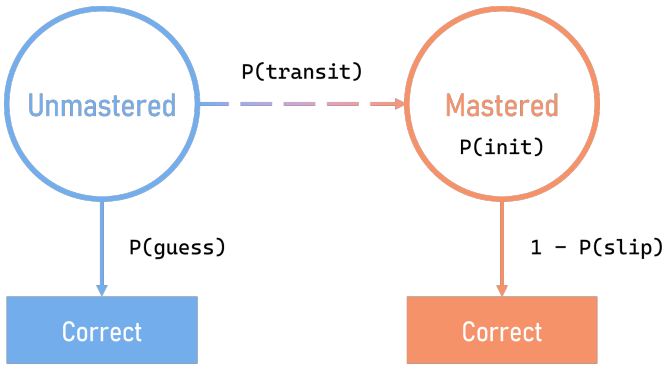
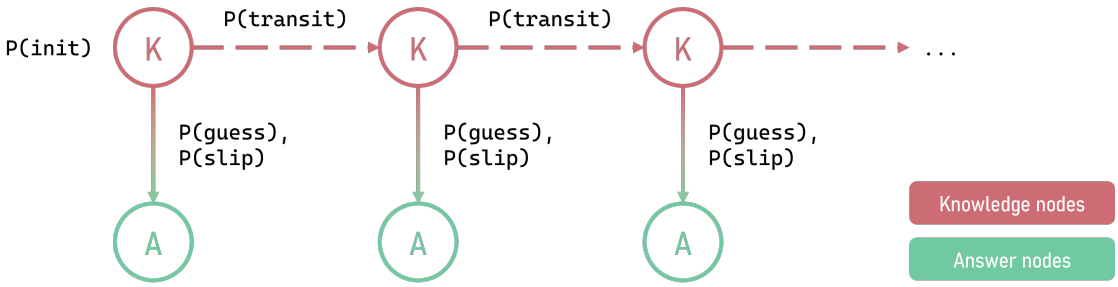


Figure 2.3: Bayesian Knowledge Tracing in Action



interpretable algorithm to explain to non-computer scientists. BKT's importance and prevalence in modern learning technologies also makes it an ideal pilot algorithm for our work at the intersection of education and AI. My thesis builds off existing evidence that verbal and visual explanations of BKT increase confidence, trust, and perceived accuracy of the algorithm [110].

2.3.1 The Algorithm

Bayesian Knowledge Tracing is an artificial intelligence algorithm that predicts student skill mastery [25]. BKT is commonly used in intelligent tutoring systems across the United States, including the *Open Analytics Research Service* [17] and *Lynette* [53], to allow for more accurate student modelling and personalized learning experiences [27, 112]. Previous research also suggests that learners prefer BKT over simpler knowledge tracing algorithms such as N-Consecutive Correct Responses (N-CCR) [110], strengthening the motivation for increasing the interpretability of BKT.

At its core, BKT can be viewed as a two-node dynamic Bayesian network, or a two-state Hidden Markov Model (HMM), that labels each skill as "mastered" or "unmastered" for a given student at a given time, based on their performance on previous practice problems [25, 53] (Figure 2.2). Observations are also binary; students either get a question right or wrong [112]. Thus, the system contains two notions of mastery: *observed* mastery as determined by the student's sequence of correct or incorrect responses (i.e., answer nodes) and *latent* mastery as determined by BKT's probabilistic

Table 2.2: Parameters of Bayesian Knowledge Tracing

Parameter	Description
P(init)	the probability that the student already knew the skill before a practice opportunity
P(transit)	the probability that the student will learn the skill after a practice opportunity
P(guess)	the probability that the student will guess correctly on an unknown skill
P(slip)	the probability that the student will make a mistake and "slip" on a known skill

estimates (i.e., knowledge nodes). These notions of mastery are illustrated in Figure 2.3. Depending on its *latent* assessment of student performance, BKT can guide the learning analytics system in selecting practice problems to complete or content to review.

As shown in Figures 2.2-2.3 and defined in Table 2.2, BKT uses four parameters to generate its estimates of skill mastery: **P(init)**, **P(transit)**, **P(guess)**, and **P(slip)**. Each of these Bayesian priors can take a value between 0 and 1. To begin, BKT sets **P(learned = 1)**, the probability of skill mastery, to the initial probability that the student learned the skill a priori, as shown in Equation 2.1. For simplicity and readability, we will use **P(learned)** to denote **P(learned = 1)** in all the equations below.

$$P(learned_1) = P(learned_0) \tag{2.1}$$

Next, the conditional probability that the student has learned the skill previously (at time $n-1$) is computed using either Equation 2.2 or 2.3 depending on whether the student answered the most recent problem (i.e., observation n) correctly. By Bayes rule, **P(obs)** is equivalent to the denominators of the following two equations.

$$P(learned_{n-1}|obs_n = corr) = \frac{P(learned_{n-1}) * (1 - P(slip))}{P(learned_{n-1}) * (1 - P(slip)) + (1 - P(learned_{n-1})) * P(guess)} \tag{2.2}$$

$$P(learned_{n-1}|obs_n = incorr) = \frac{P(learned_{n-1}) * P(slip)}{P(learned_{n-1}) * P(slip) + (1 - P(learned_{n-1})) * (1 - P(guess))} \tag{2.3}$$

This conditional probability is then used to update the probability of skill mastery (at time n) as shown in Equation 2.4.

$$P(learned_n|obs_n) = P(learned_{n-1}|obs_n) + (1 - P(learned_{n-1}|obs_n)) * P(transit) \tag{2.4}$$

We update **P(init)** with the new value of **P(learned)**, repeating this process as the student completes additional practice problems. Figures 2.2 and 2.3 show this process in action. A **P(init)** value of 0.95 is typically used as the qualification threshold for skill mastery [27], but this threshold criterion can vary depending on the application of BKT (e.g., middle school vocabulary quiz vs. medical school entrance exam). The other priors, **P(transit)**, **P(guess)**, and **P(slip)**, are typically not updated throughout the learning exercises, remaining at their pre-set initial values [25].

In practice, these parameters are fit through a variety of methods, most often using data from a previous offering of the same class or from students who used the skill of interest without an intelligent tutor [27]. If previous student data is not available, subject matter experts (SMEs) may be asked to determine appropriate starting parameter values to address the "cold start" problem [93]. Alternatively, Expectation-Maximization (EM) or brute force grid search search methods could be used to fit parameters for BKT [49]. Some limitations of these parameter fitting methods are discussed in the following section.

2.3.2 Limitations of BKT

Since BKT is a probabilistic algorithm, it falls subject to certain biases and limitations. For example, BKT is vulnerable to **model degeneracy**, which occurs when the algorithm does not work as expected due to its initial parameter values being outside of an acceptable range [33]. This is why **P(slip)** is typically bounded between 0 and 0.1, while **P(guess)** is bounded between 0 and 0.3; otherwise, BKT may predict that a student is more likely to gain mastery by answering a question incorrectly, rather than correctly [27]. Other conditions necessary for non-degeneracy include that **P(init)** cannot be 0 or 1, **P(transit)** cannot be 1, **P(guess)** cannot be 1 - **P(slip)**, and each student must have at least two observations [32]. If these conditions are not met, multiple parameter settings can fit the same data equally well [3], meaning highly divergent **P(guess)** and **P(slip)** values may produce identical performance and result in non-identifiable models.

Different numbers of observations, n, can also lead to various forms of model degeneracy. While a low n often produces high **P(slip)** values, a high n often produces high **P(guess)** values [32]. If EM is used to fit parameters, there is the potential problem of getting stuck at local minima as well since BKT corresponds to a non-convex optimization problem [30, 57]. This may prevent the system from finding the true global optimum [93]. EM is also not guaranteed to converge for latent data models due to singularities [30, 41]. Although the problems of getting stuck at a local minima and non-convergence are two separate issues, they can often co-occur.

Additionally, BKT parameters are often shared across an entire class of students [33, 81, 112], leading to inequitable outcomes due to lack of individualization, particularly in terms of student learning rates. However, BKT's assumption that learning is all-or-nothing leads to inherent model misspecification, so even when different parameters are fit for fast and slow learners, the problem of inequity persists [33]. In the true student model, learning happens much more often and in more progressive increments than BKT makes it appear, but introducing additional levels of mastery can lead to degenerate parameter estimates as well [32].

The BKT parameters also do not account for certain events, such as forgetting [32] or the time it takes a student to answer a question [45], which are relevant and important to consider when assessing learning and mastery. Furthermore, BKT assumes all skills are independent and that questions of the same skill do not differ in difficulty [29].

Without a proper mental model for BKT and its limitations, students (and teachers) may make decisions based on their own observations of a learning analytics system's outputs, which may not be accurate. This is where XAI can help. A sufficient understanding of the underlying algorithm

could influence trust in the system as well as student decision-making for learning, thus forming the motivation for this work.

2.3.3 BKT Variants

In our study, we focus on the standard BKT model introduced in Section 2.3.1. However, several "modified" versions of BKT have been constructed over the years, each targeting various subsets of the limitations detailed above to increase the algorithm's wider efficacy and applicability. Some variants target student-level parameter individualization [112]. Others, like *KT-IDEM (Item Difficulty Effect Model)*, allow for differing values of **P(slip)** and **P(guess)** depending on the question type to take item difficult into account [82]. BKT models have also been incorporated into Massive Open Online Courses (MOOCs) to customize parameters based on educational resource type (e.g., videos vs. tutorial vs. discussion board) [80]. Additionally, some BKT extensions add a parameter for forgetting [90] and consider the impact of response time on model predictions [45]. Other variations of BKT are discussed in [11, 83].

2.4 Summary

In this chapter, I introduce explainable AI as a way for increasing algorithmic transparency and instilling a more realistic sense of trust into users of AI-mediated systems, focusing on the potential benefits and applications of post-hoc explainables. I also motivate the need for applying methods from the learning and cognitive sciences to better understand what people need to know about AI algorithms and ultimately construct better post-hoc AI explanations. Specifically, I discuss how Cognitive Task Analysis can be used to identify the necessary knowledge components to teach stakeholders about a topic, such as an AI algorithm, and how these KCs can be used to design assessments and explainables to teach the algorithm in question. I end the chapter by providing an overview of Bayesian Knowledge Tracing, an AI algorithm commonly used to predict mastery in learning analytics system. I also share some of the limitations and variants of BKT, and elucidate the reasons why we selected it as the pilot algorithm for my thesis.

Chapter 3

Prior Work

This chapter focuses on describing relevant work completed prior to my thesis toward creating a more robust, evidence-based framework for XAI [111]. Much of this work was done in conjunction with Noah Cowit '20 and aligns with the first three steps in our proposed framework (Figure 1.2). First, I will describe the methodology we used to conduct expert Cognitive Task Analyses (CTA) of Bayesian Knowledge Tracing (BKT) to gain general knowledge about expert understanding with respect to our selected AI algorithm. Then, I will present the resultant knowledge components derived from CTA that were targeted in this work when designing a post-hoc explainable for BKT.

3.1 Cognitive Task Analysis for BKT

To gain a deeper understanding of how experts approach and reason about problems pertaining to BKT, we conducted CTAs with student experts of the algorithm. Our CTA protocol essentially involved interviewing these students and having them step through various scenarios that may be encountered when using a BKT system while thinking aloud [103]. This is analogous to the approach described in [68]. Our participants were seven undergraduate students at Williams College who had previously studied BKT as part of past research experiences and in some cases had implemented small-scale BKT systems themselves. Interviews were semi-structured with a central focus on responding to a script of ten problems and lasted between 30-60 minutes in duration. By recording comprehensive, qualitative information about expert performance during these interviews as in [24], we were able to identify the knowledge components of BKT expertise.

3.1.1 Vignette Surveys

For the context of this study, we developed our own CTA problems, as identifying problems for BKT experts to solve is less straightforward than identifying problems for statistics experts to solve for example, as in [68]. To generate problems that experts would realistically encounter when interacting with BKT systems, we adapted our approach from **Vignette Survey** design [10]. Vignette Surveys typically involve using short, descriptive scenarios to obtain individual feedback and are especially

Table 3.1: Example CTA Problem Probing Basic Comprehension of P(slip)

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. 2. Amari got 6 out of 10 questions correct on their test. At the end of the test, BKT suggests Amari has not mastered the material. What is your interpretation of this analysis?

helpful for situations where the number of testable characteristic parameters is large [10]. This is the case for BKT due to the continuous nature of the algorithm's four key parameters (i.e., **P(init/transit/guess/slip)**). Even if we simply looked at low, medium, and high values for each parameter, this would still result in $4 \times 3 = 12$ conditions, which is not insubstantial when we consider the time, resources, and number of participants required to run a traditional within- or between-subjects experiment.

Additionally, the numerous social indicators of BKT parameters and weighing of subjective factors necessary for model evaluation make vignettes an effective 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)**. Previous work has also used similar vignette-style approaches to evaluate fairness by presenting participants with hypothetical scenarios involving algorithmic and human managerial decisions [63].

Each scenario posed to our experts involved a vignette describing background information about a hypothetical student followed by one or more questions regarding BKT. One such scenario relating to **P(slip)** included in our CTA protocol is presented in Table 3.1. A sample expert response to this problem is shown in Table 3.2. The numerical bullets under the topic headings correspond to our resultant knowledge components of BKT, as discussed in Section 3.2.

We were not only interested in expert comprehension of the parameters and equations involved in BKT, but also the context in which BKT systems are used. Thus, Table 3.3 shows another sample question from our CTA protocol that explores expert understanding of the robustness of mastery measures computed by BKT systems. Additional problems used in our CTA are included throughout this thesis; for the full list, please see Appendix A.

3.2 Knowledge Components of BKT

To derive knowledge components (KCs) from interview transcripts, we first identified the initial and final states (i.e., the given information and goal) of each scenario. Questions with similar objectives were grouped together, forming larger knowledge areas (e.g., "Identifying Priors"). Next, each

Table 3.2: Example Expert Interview Excerpt Corresponding to Table 3.1

Evaluating P(init):

1 Amari does know the words, even though they get mixed up sometimes,

2/3 but they still have the knowledge,

4 so maybe I'll say 0.6.

Identifying Changed Parameters:

1/4 P(transit) might be lower than the last scenario (0.7) because

2 you don't get feedback right away,

3 so it would be harder to get improvement, and improve on the next question.

5 I'll do like 0.5.

Identifying Priors:

5 P(guess) is 0.1

1/2 because there are 10 questions.

Identifying Priors:

4 P(slip) seems relatively high because

2/3 Amari gets flustered and mixes up words they would otherwise get correct, so it's not supposed to be higher than 0.1,

5 but I'll do 0.3.

Limitations of BKT:

1 Based on BKT this means P(init) was not 0.95 or higher.

3 But in real life, Amari may know all the words.

4 Because Amari mixes up the words when they have to take a test, it may not be accurate what BKT says.

Table 3.3: Example CTA Problem Probing Real-World Use of BKT

Kim recently took an extensive test on Nuclear Powerplant operations. The test was a standard 4 question multiple choice exam, and Kim finished in a reasonable timeframe. After taking the test, Kim's P(Init) is 0.97. Kim lives near a nuclear power plant, and the operator is out sick for the day. The powerplant desperately needs a temporary operator, and would rather not pay to helicopter one in. If the operator is not a master in Nuclear Powerplant operations, a nuclear meltdown is likely to occur. Should Kim be offered the job? Explain your response.

Table 3.4: Knowledge Components for Identifying Priors

Identifying Priors: 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 the limitations of BKT and how this could impact the value of this parameter.* 4. Make an assessment about the parameter in question based on this qualitative evidence (or lack thereof). 5. Choose a value for the parameter by converting assessment to a probability between 0 and 1.

participant's responses were coded to identify the steps taken to achieve the goal from the initial state. Then, we noted the common steps that participants used in each scenario. Final KCs 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 in the following subsections by using *italics*.

We ultimately divided our analysis of BKT into four discrete but related knowledge areas: **Identifying Priors**, **Identifying Changed Parameters**, **Evaluating P(init)**, and **Limitations of BKT**. Each knowledge area consisted of 4-5 knowledge components, resulting in a total of 19 KCs across all knowledge areas [111].

3.2.1 Identifying Priors

The **Identifying Priors** knowledge area concerns the processing of subjective vignette material into reasonable numerical values for the four initial parameters of BKT. We identified five basic components (including one optional KC) of expert processes in formulating this knowledge, which were almost identical for each Bayesian prior and are listed in Table 3.4. Question 1 in Table 3.1 shows an example problem categorized under this knowledge area.

Table 3.5: Knowledge Components for Identifying Changed Parameters

Identifying Changed Parameters:

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.

3.2.2 Identifying Changed Parameters

The **Identifying Changed Parameters** knowledge area involves a very similar cognitive process to Identifying Priors. The major difference is that in the case of identifying initial parameters, the expert ultimately arrives at a numerical value (e.g., 0.7), while in the case of identifying changed parameters, the direction of change (e.g., increase), if any, is the most important factor. Participants were not necessarily asked to produce a numerical value after making this direction assessment. We identified five KCs in this knowledge area, as shown in Table 3.5.

An example question categorized under Identifying Changed Parameters is: "Sandy gets the first answer correct. What are reasonable values for the BKT parameters now? Please talk me through your reasoning" (Question 1b, Appendix A). Problems in this knowledge area often followed those classified as Identifying Priors.

3.2.3 Evaluating P(init)

The **Evaluating P(init)** knowledge area addresses the fact that understanding **P(init)** in particular is essential for evaluating real-life, practical applications of BKT. As previously discussed, in standard practice, a value of **P(init)** over 0.95 is considered "mastery" [27]; however, this threshold can vary. An analysis of **P(init)** that takes into account the probabilistic nature of the BKT formula indicates a more nuanced understanding of the algorithm.

For example, the question in Table 3.3 targets this more nuanced understanding. All of our experts were hesitant to label Kim as a "master" of nuclear powerplant knowledge and offer her the temporary job, even though she received a **P(init)** of 0.97 on a comprehensive test scored by BKT. For example, one participant stated, "Even though [Kim] has a pretty high P(init), so she has the theoretical knowledge, but she may not have the necessary applied knowledge for this job." Another participant explained, "It's just like such a high stakes job that you want someone with experience and actual know-how... that is just too much of a human stake to put on an algorithm in my mind."

Table 3.6: Knowledge Components for Evaluating P(init)

Evaluating P(init):

1. Synthesize information from vignette, considering the parameter level of **P(init)** in particular.
2. Make a judgment as to the magnitude of **P(init)** (e.g., low, moderate, high, moderately high, etc.).
3. Consider this magnitude with respect to the situation described in the vignette and BKT's definition of mastery. Some situations call for a very high level of knowledge—and thus a very high **P(init)**—to be considered mastery (e.g., medical tests, 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 (usually by answering yes/no or making a final judgment concerning mastery). Questions often took the form of "With this value of **P(init)**, has X achieved mastery?" or "Is 0.4 a reasonable value for **P(init)** in this situation?"
5. *Explain why BKT's predictions might not be accurate in this case due to its limitations, probabilistic nature, etc.*

This observed thought process was relatively standard across questions that required using **P(init)** to evaluate mastery, resulting in five components of understanding (including one optional KC) illustrated in Table 3.6. These questions are more open-ended in nature, but our participants typically followed similar paths to arrive at their respective conclusions.

3.2.4 Limitations of BKT

In addition to asking about algorithmic understanding, we also expected experts to have a sophisticated understanding of BKT's strengths and weaknesses, thus prompting the **Limitations of BKT** knowledge area. As a whole, these questions interrogated the discrepancy between what BKT is able to predict and what an ideal knowledge tracing algorithm would be able to predict.

We covered three limitations of BKT within this protocol. The first was model degeneracy, where BKT exhibits unexpected behavior due to invalid parameter values, as defined by [33]. Question 2 in Table 3.1 illustrates an example problem where participants grappled with a degenerate model, as Amari's flustered exam state likely translates to a high **P(slip)**. We also tested experts on additional parameters that BKT does not currently track, such as time taken to complete an exam and forgetfulness between tests; a sample problem targeting this latter limitation is shown in Table 3.8. Finally, we asked our experts to consider the probabilistic nature of BKT and how the algorithm's inherent uncertainty may limit its reliability in high stakes scenarios, as in the nuclear power plant example in Table 3.3. In many cases, problems involving Limitations of BKT were also related to the Evaluating P(init) knowledge area.

Again, these problems tended to be more open-ended and complex, similar to those categorized under Evaluating P(init). As a result, participant dialogue was often more oscillatory than methodological, a likely reflection of the uncertain cognitive processes they were experiencing while

Table 3.7: Knowledge Components for Limitations of BKT

Limitations of BKT:

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.

Table 3.8: Example CTA Problem Probing the Limitation of Forgetting

Marc has just returned to school after a summer vacation. Her teacher gives her a handout with review problems. Marc cannot remember how to add fractions with different denominators for her life. Marc is upset. She got a very good grade on the test for the material last year.

1. If this previous test were evaluated by BKT, what would be a reasonable value for P(Init) before Marc attempts the review problems?
2. Describe this behavior of BKT and why it is or is not intuitive/sensible.

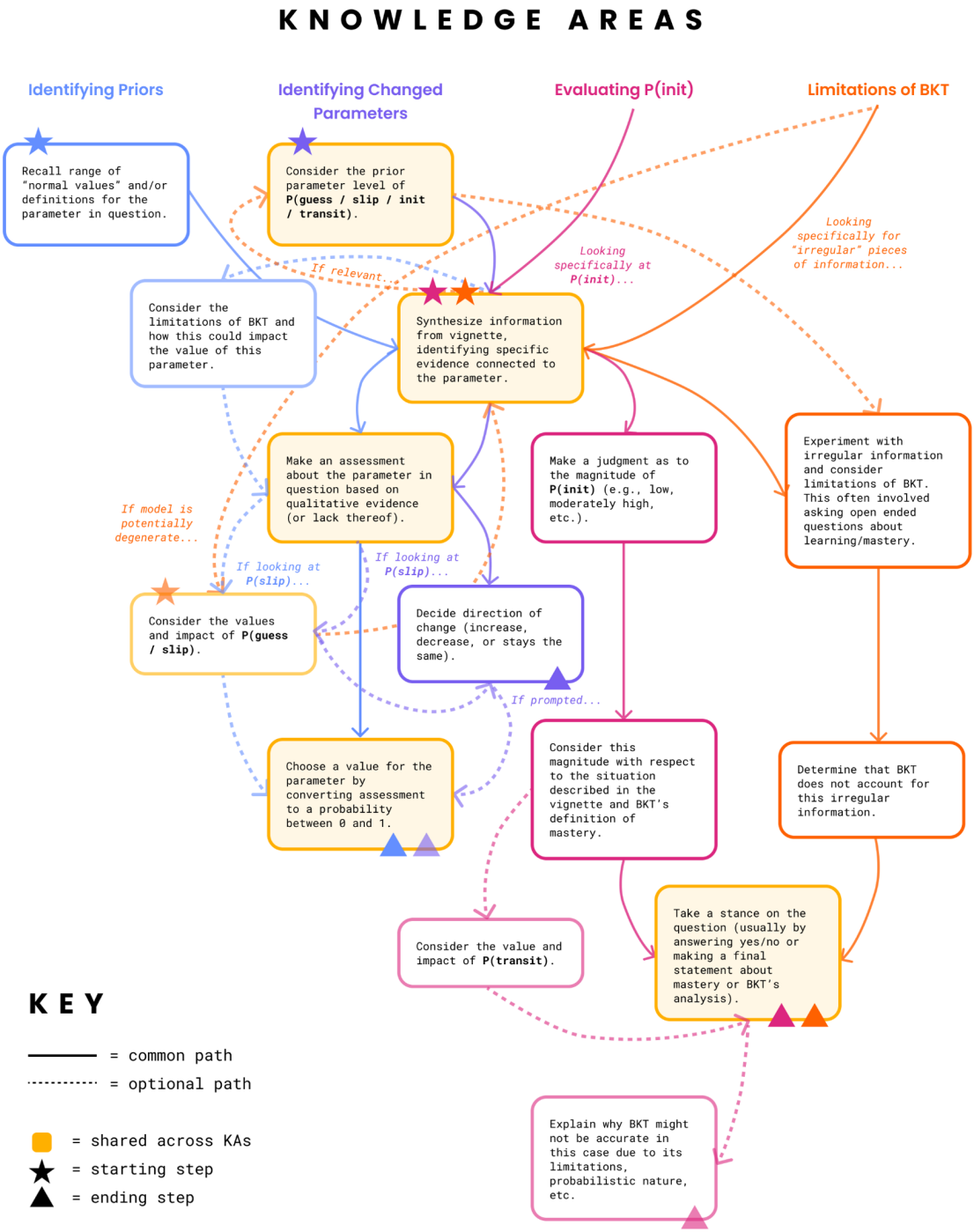
thinking through the situations presented. We identified four components (including one optional KC) of expert processes in this knowledge area, detailed in Table 3.7.

3.3 Visualization of KCs

To better visualize how our resultant knowledge components and areas for BKT intersect and interact, I designed an interactive flow chart[1] using JavaScript/jQuery (version 3.5.1), HTML, and CSS [111]. A screenshot is included in Figure 3.1 for reference. In my flow chart, knowledge areas are delineated with different colors, and arrows are used to show the flow of KCs (i.e., the order of steps used by participants to reason about our CTA scenarios). Common paths taken by BKT experts are denoted with solid lines (i.e., those taken by the majority of participants), while optional paths are denoted with dotted lines. Additionally, starting steps are marked with a star symbol, while ending steps are marked with a triangle symbol. Any shared KCs between our four knowledge areas are filled in with light orange.

[1] https://catherinesyeh.github.io/bkt-kcs/

Figure 3.1: Flowchart Depicting Knowledge Components and Areas of BKT



In particular, this flow chart helped us visually capture the common thought processes that participants used across problems spanning different knowledge areas of BKT. For example, the KC: "Synthesize information from vignette, identifying specific evidence connected to the parameter" served as a starting step for both **Evaluating P(init)** and **Limitations of BKT**. Similarly, the KC: "Choose a value for the parameter by converting assessment to a probability between 0 and 1" was a potential ending step for questions categorized under the **Identifying Priors** and **Identifying Changed Parameters** knowledge areas.

We made some minor modifications to our list of KCs after further analysis of interview transcripts, which is why there may be some discrepancies between this visualization and the KCs described above in Section 3.2. However, the broader knowledge areas and general flow/ordering of KCs should still be the same.

3.4 Summary

In this chapter, I discuss the preliminary work I completed prior to this year, which serves as a foundation for my thesis. For example, we applied Cognitive Task Analysis to Bayesian Knowledge Tracing, which involved interviewing student BKT experts and asking them to think aloud while stepping through various vignette-style scenarios pertaining to BKT systems. Our CTA revealed four larger knowledge areas of BKT: (1) **Identifying Priors**, (2) **Identifying Changed Parameters**, (3) **Evaluating P(init)**, and (4) **Limitations of BKT**. We then broke these knowledge areas down further to form 19 smaller knowledge components that represent the common steps taken by experts to solve our CTA problems and ultimately the necessary concepts that should be taught to BKT stakeholders. I also present a interactive flowchart I made to better visualize the interaction between our BKT knowledge areas and KCs.

Chapter 4

Methodology

In this chapter, I will elucidate the methodologies and processes I used during my thesis to create and evaluate my explainable for Bayesian Knowledge Tracing. This work follows directly from and serves to address the results obtained previously via Cognitive Task Analysis (Chapter 3), mirroring our proposed XAI framework as illustrated in Figure 1.2.

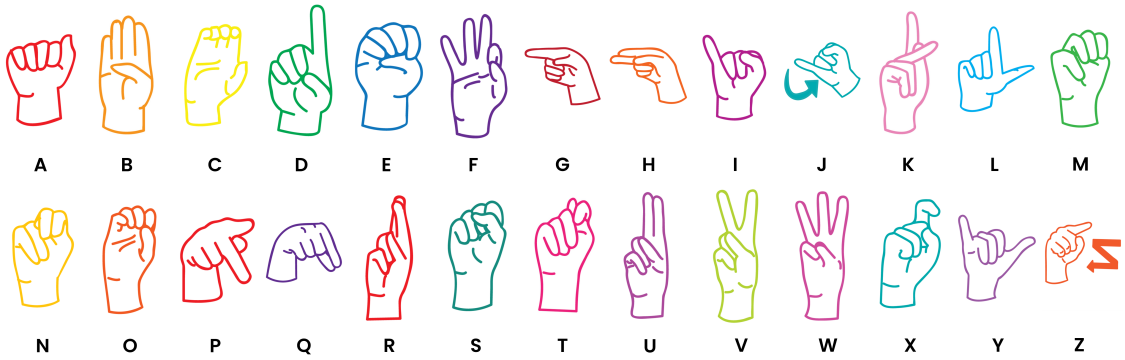
Immediately after determining the KCs of BKT through CTA, I designed assessments targeting these KCs as per Backward Design [109], which will be introduced when discussing evaluation in Section 4.3. Next, I designed and implemented my explainable using an iterative design process [31, 62]. Finally, I conducted user studies to assess my final BKT explanation. The resultant explainable is an interactive web application that uses American Sign Language to motivate and illustrate the behavior of BKT systems.

4.1 Explainable Design

Following the iterative methodology [31, 62] depicted in Figure 4.1, my explainable design process involved several cycles of brainstorming, prototyping, testing, and revising. This approach was inspired by principles from user-centered design—in particular, the five design thinking modes from Stanford's d.school (i.e., the Hasso Plattner Institute of Design): empathize, define, ideate, prototype, and evaluate [31].

Figure 4.1: The Five-Step Design Process

Figure 4.2: The Fingerspelled Alphabet for American Sign Language



4.1.1 Brainstorming

After completing my preliminary literature review (Chapter 2) and determining the KCs of BKT using CTA (Chapter 3), which comprised the *Empathize* and *Define* phases of the design process (Figure 4.1), I moved on to the *Ideate* stage.

While brainstorming, we first considered what type of explainable to build and how it would be structured. Previous ideations of BKT explainables had been largely metaphor-based, using familiar concepts such as hot air balloons[1] or apple-picking[2] to illustrate and teach the more complex aspects of BKT. However, this time, we decided to try a different approach, centering our explainable around teaching an actual skill to participants to demonstrate the real-world applicability of BKT. This skill-based approach also allows us to show BKT in action on a smaller scale and provides participants with the opportunity to learn something new through interacting with our explainable.

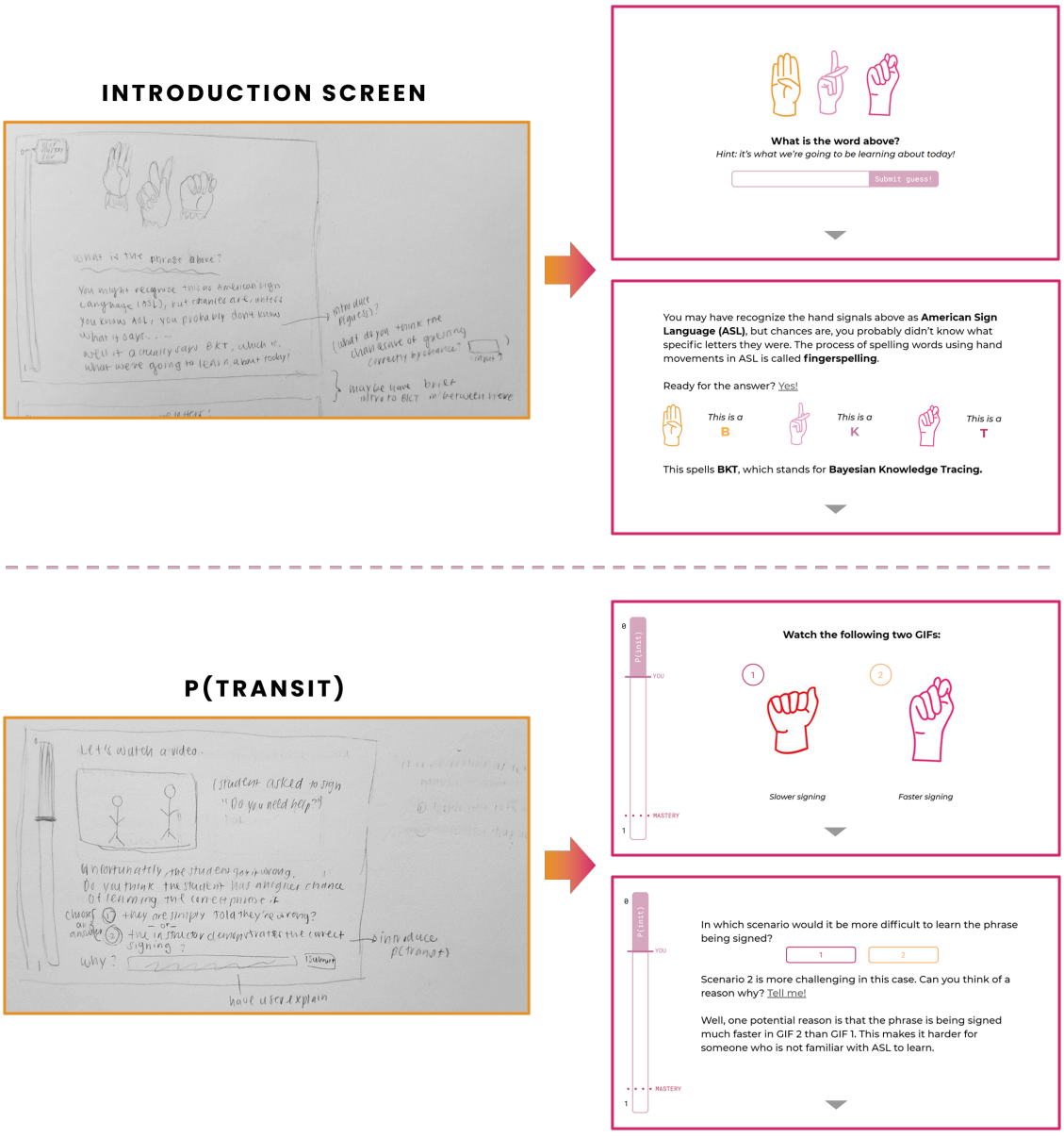
American Sign Language

The skill I ended up selecting for my explainable was **American Sign Language (ASL)**, the primary language and form of communication for deaf or hard-of-hearing people in North America [75]. We chose ASL because we believe it is a useful, valuable skill to learn, and it is still relatively uncommon, especially among hearing people. As of 2011, less than 1% of the US population knew ASL [15]. Sign languages are also distinct from both written and spoken languages, leaving them unsupported by the majority of modern communication technologies [37].

One key component of ASL is **fingerspelling**, which is the process of spelling out words using hand shapes corresponding to the letters in the word [75]. The American fingerspelled alphabet is displayed in Figure 4.2 and more examples from our explainable are shown in Figures 4.3, 4.4, 4.7, and 4.8. All ASL alphabet artwork used throughout my prototypes and final design was created by Darlene Aalbert[3]. To keep things on the simpler side for our participants, as this project is meant to serve as an introduction to ASL, I decided to limit my explainable to teaching the fingerspelled

[1] https://catherinesyeh.github.io/bkt-balloon/
[2] https://minhbphan.github.io/ApplePickingBKT/
[3] https://darlenealbert.myportfolio.com/

Figure 4.3: Evolution of BKT Explainable Prototypes



alphabet and applying it to spell words. In general, however, fingerspelling is primarily used to indicate proper nouns or other words/phrases for which there is no official sign in ASL [75].

4.1.2 Lo-fi & Hi-fi Prototypes

Next, I proceeded to the *Prototype* stage in the design process (Figure 4.1), which involved physically sketching and rendering a rough draft of my BKT explainable. To begin, I completed a few low-fidelity (lo-fi) prototypes [91] to get a sense of how I wanted to incorporate ASL and organize the explainable. After some modifications and additional brainstorming sessions, I translated these lo-fi sketches to an interactive, high fidelity (hi-fi) prototype—essentially a digitalized paper prototype [96]–in Google Slides. This higher fidelity prototype detailed nearly all of the design decisions to be implemented during the explainable implementation phase. Some snapshots of my initial lo-fi sketches and their corresponding hi-fi screens are shown in Figure 4.3. I was originally considering using Figma to create my hi-fi prototype as well, but eventually settled on Google Slides due to familiarity and ease of use—both for me, the researcher, and the participants.

Content Design

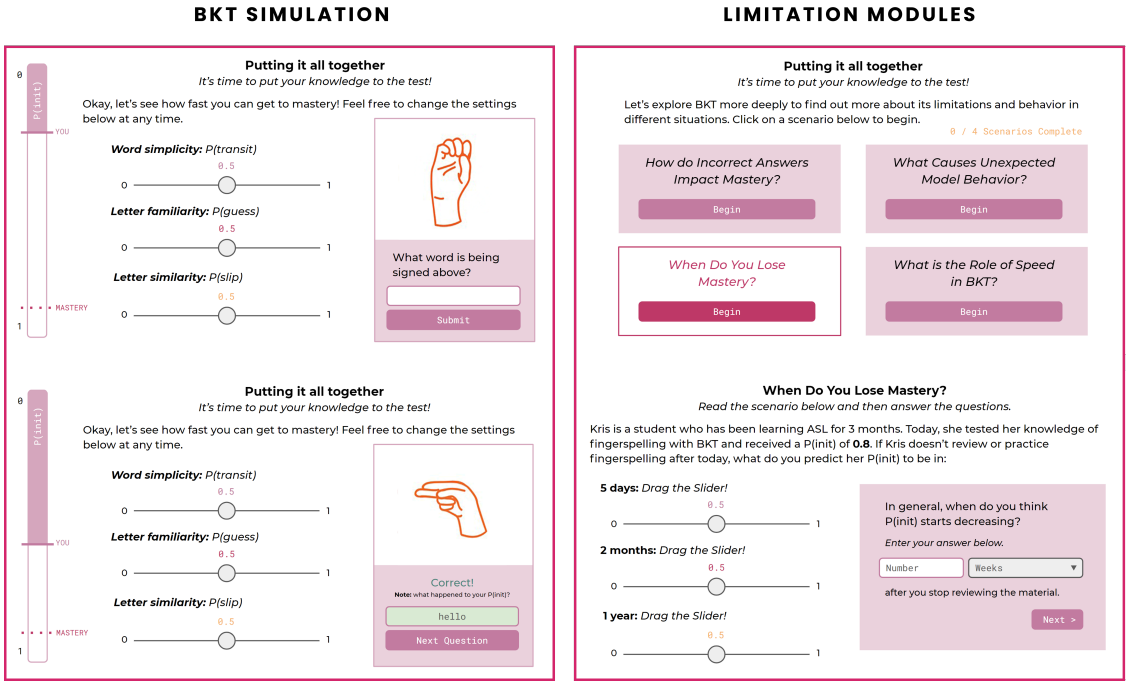
To ensure that our explainable design was effective and conducive to learning, we made sure to follow best practices in pedagogy, focusing particularly on *active learning* and *self-explanation*. Interactive activities that involve "learning by doing" result in greater learning than passively intaking information through reading or watching videos [60]; this idea is embodied throughout my explainable by asking the user to choose an answer response or type an open-ended response in order to complete various activities.

Hake defines interactive engagement as "those [activities] designed at least in part to promote conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors." [47]. Thus, simply scrolling through a webpage or clicking a "Next" button is not considered active learning by Hake's standards. However, it should be noted that pedagogical design and interaction design are two separate sets of constraints to follow. While navigational elements such as scroll bars and buttons may not serve much purpose for learning, they are important to supporting interface functionality.

Throughout the explainable, we complemented required opportunities to provide a "heads-on" response [56] with additional prompts for the user to consider their answer to particular questions without a provided answer box, encouraging mental engagement and inquiry-based learning [39]. I also aimed to incorporate *desirable difficulties*, which involves introducing selective challenges and cognitive burden to the user in order to promote deeper engagement and analytical reasoning [21]. Strengthening the claim of active learning over passive consumption, research shows that if a student provides an incorrect response to a question, as long as the test is followed by immediate corrective feedback, this will lead to greater learning than a single presentation of the correct answer [50].

We embedded this idea in our design by providing immediate feedback to the user as they progress through the explainable so that even in cases where we cannot directly evaluate the user's response,

Figure 4.4: Additional Snapshots of Hi-fi Explainable Prototype



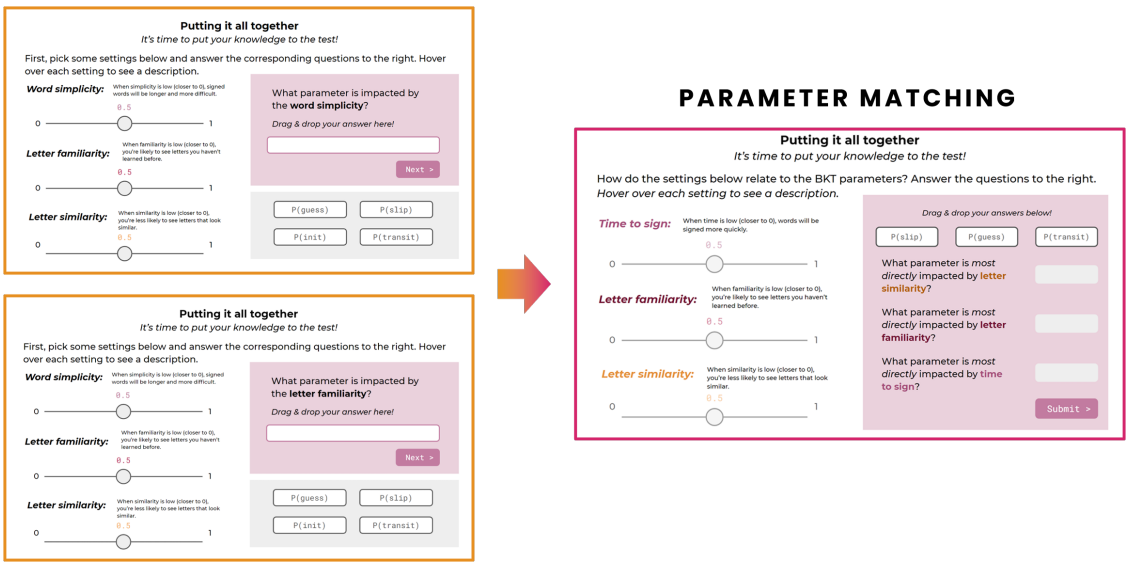
we can offer some form of guidance to aid their learning. Furthermore, researchers have established that when people explain examples to themselves, they learn more [104]. The most effective learners will make more comments about the conditions under which specific actions were preferred, the relationships between actions and goals, and the consequences of different actions [104].

I implemented these learning theory principles in my explainable by first introducing each of BKT's four key parameters separately using various self-explanation prompts [102] and interactive activities (Figure 4.3), which require the user to make predictions about content pertaining to ASL or BKT. Then, I bring the parameters together, culminating in a mini game/simulation to show BKT in action (Figure 4.4). This organizational structure was inspired by another previous BKT explainable[4]. Participants are also able to track their own learning progress throughout the explainable via the main mastery bar (see Figure 4.3 bottom right and Figure 4.4 left).

After the BKT mini game, I also added four modules to teach BKT's flaws and limitations (Figure 4.4). This is essential to our goal of encouraging deeper exploration of the algorithm and helping users develop realistic trust in BKT systems but something our earlier explainables did not explicitly touch on. To ensure that we were targeting the BKT KCs identified previously in Section 3.2, we mapped each question/activity in our prototype explainable to its corresponding KC(s), following the approach outlined in [20]. For example, the activities in our "When Do You Lose Mastery?" limitation module (Figure 4.4, right), which involve making assessments about the magnitude of **P(init)**, directly map to KC 3 under the **Identifying Priors** knowledge area (Table

[4]http://www.cs.williams.edu/ iris/res/bkt-esperanto/

Figure 4.5: Redesign of Parameter Matching Activity



3.4) and KC 2 under the **Evaluating P(init)** knowledge area (Table 3.6). See Appendix C.1 for the full mapping of explainable activities to KCs.

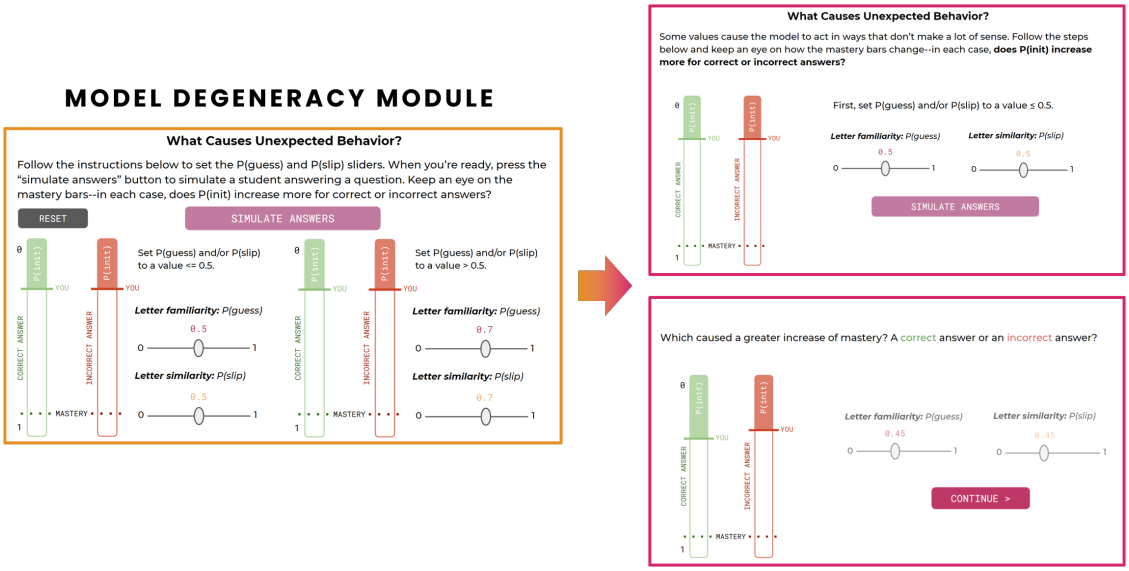
4.1.3 Usability Testing

User Studies

After the hi-fi prototype was complete, I conducted a series of **pilot user studies** with the assistance of Mira Sneirson '22 to evaluate usability and improve my explainable design—this brought us to the *Evaluate* phase of the design process (Figure 4.1). Poorly-designed interfaces can negatively impact learning outcomes [8], so usability testing was key to ensuring that my design was effective and conducive to learning. User testing can also help identify problems with a design that may not obvious to the researchers themselves [91, 96]. During these user studies, I asked participants to think-aloud while stepping through each slide and interacting with my prototype to complete the necessary tasks, mirroring traditional usability evaluation sessions [96]. In total, five Williams CS students were asked to participate in this initial round of user studies.

The participant feedback we received from user testing was used to further hone and modify my hi-fi prototype. For example, we noticed students getting stuck on the parameter matching activity preceding our mini BKT simulation. During our user studies, it became apparent that some participants were unclear how many parameters could be selected for each question and reused the same answer choices multiple times. Others also wondered if they were supposed to answer the questions by dragging the sliders on the left. To address these issues, I made the sliders inactive and restructured the matching questions on the right to clarify that each parameter should only be used once. I also adjusted some of the wording on the screen and renamed the top parameter from "word simplicity" to "time to sign" in order to make the connection between ASL and BKT more

Figure 4.6: Redesign of Model Degeneracy Module



evident and logical for participants. These changes are illustrated in Figure 4.5, with my original design outlined in orange, and my new design outlined in pink.

Similarly, several participants pointed out how our original model degeneracy module was excessively confusing and likely a source of cognitive overload [87] due to the overwhelming amount of information and objects packed onto the screen. Cognitive overload was also identified as a common issue in XAI systems by previous work looking at designing explanations for reinforcement learning algorithms [6]. Introducing desirable difficulties [21] without inducing cognitive overload or other harmful drawbacks of increased difficulty is a known problem in the learning sciences as well [79]. This prompted several rounds of redesign, and ultimately we decided to separate the module into multiple, simplified screens, as shown in Figure 4.6, to improve the overall user experience. Once again, my original design is outlined in orange, and my new design outlined in pink. Here, I also made sure to deactivate the sliders when appropriate, so participants would not be confused when they were supposed to adjust the parameter values. Our later user study subjects, who represented an independent sample of users from the first, described this new design as more straightforward and approachable, and they had a much better sense of what the module was asking them to do.

I also made some other minor changes to my explainable after conducting usability evaluations (e.g., wording, images, colors, etc.), but the two cases described above and depicted in Figures 4.5-4.6 regarding the parameter matching activity and model degeneracy module were the most significant updates completed during this stage of the design process.

Heuristic Evaluation

In conjunction with our user studies, one additional step we took to evaluate the usability of my explainable prototype before implementation was **heuristic evaluation** [77]. Our heuristic eval-

Table 4.1: Nielson's Ten Usability Heuristics

#	Heuristic	Description
1	**Visibility**	Show system status & tell what's happening
2	**Real World Mapping**	Use familiar metaphors & language
3	**Control & Freedom**	Provide good defaults & undo
4	**Consistency**	Use same interface & language throughout
5	**Error Prevention**	Help users avoid making mistakes
6	**Recognition (not recall)**	Make information easy to discover
7	**Flexibility & Efficiency**	Make advanced tasks fluid & efficient
8	**Minimalism**	Provide only necessary information in an elegant way
9	**Error Recovery**	Help users recognize, diagnose & recover from errors
10	**Help**	Use proactive & in-place hints to guide users

uation involved asking three research assistants (myself & two other students working with Prof. Howley) to examine the explainable's user interface and assessing its compliance with Jakob Nielson's ten recognized usability principles [76] (e.g., visibility, real world mapping, control & freedom, consistency, etc.). A color-coded and lightly modified version of these heuristics—courtesy of Scott Klemmer and Janaki Kumar—is included in Table 4.1.

To complete the heuristic evaluation, each evaluator stepped through the explainable, making note of each heuristic broken, rating its severity (on a scale from 1, least severe, to 5, most severe), and providing a more specific description of the problem at hand. Sample violated usability principles identified via heuristic evaluation are shown in Table 4.2, along with their user-crafted descriptions and severity rankings. Some of these heuristics could not be addressed until the explainable was implemented (e.g., heuristics 9 and 10). Others supported our findings from the user studies such as heuristic 6, which suggested simplifying the parameter matching activity (Figure 4.5) so that recall would be less challenging. Additionally, I ended up adding to my earlier description of **P(init)** on previous slides because some participants mentioned that it was unclear how **P(init)** was different from the other three BKT parameters and why it was not a potential answer option.

Heuristics like 3 prompted deeper design considerations, as adding a back button would allow users to refer back to previous screens while working through the explainable, but we were unsure whether this would be the desired behavior for participants. Thus, we left these decisions for the implementation phase as well.

Cognitive Walkthrough

We also considered conducting **cognitive walkthroughs** to inspect the usability of my BKT explainable prototype. Cognitive walkthroughs are similar to heuristic evaluations, albeit more task-based [70]. During a cognitive walkthrough, experts define tasks to complete with a user interface

Table 4.2: Selected Broken Heuristics for BKT Explainable

Heuristic Broken	Description	Severity (1-5)
3	No back button/arrow to return to previous state; there is the consistent ↓ to move forward, but no ↑	4
6	"Putting it all together" section tests users' mastery on material by answering questions based on recall (users had to have remembered **P(slip)**, **P(guess)**, and **P(transit)** for answering questions); P(init) not explained before tested?	3
9	No error messages seen (though understandable given it is a prototype)	2
10	No help and documentation seen (understandable since it is a prototype; may not be entirely necessary either)	1

and then step through each task as the user to verify if it is possible to successfully achieve each step using the proposed design [88].

However, we realized that unlike many traditional user interfaces, there were not really any clearly definable tasks that participants could complete using our explainable, besides than the general goal of stepping through the entire BKT explanation and learning about BKT. Another issue is that when learning is involved, users may not know what goals they are trying to achieve, or what steps need to be completed to reach these goals. For example, during the **P(transit)** module (Figure 4.3, bottom right), participants are likely unaware that their goal is to learn about the parameter **P(transit)**. And again, there really is not a clear task to assign users in this case, other than scrolling down the page, engaging with the provided information, and answering the posed questions.

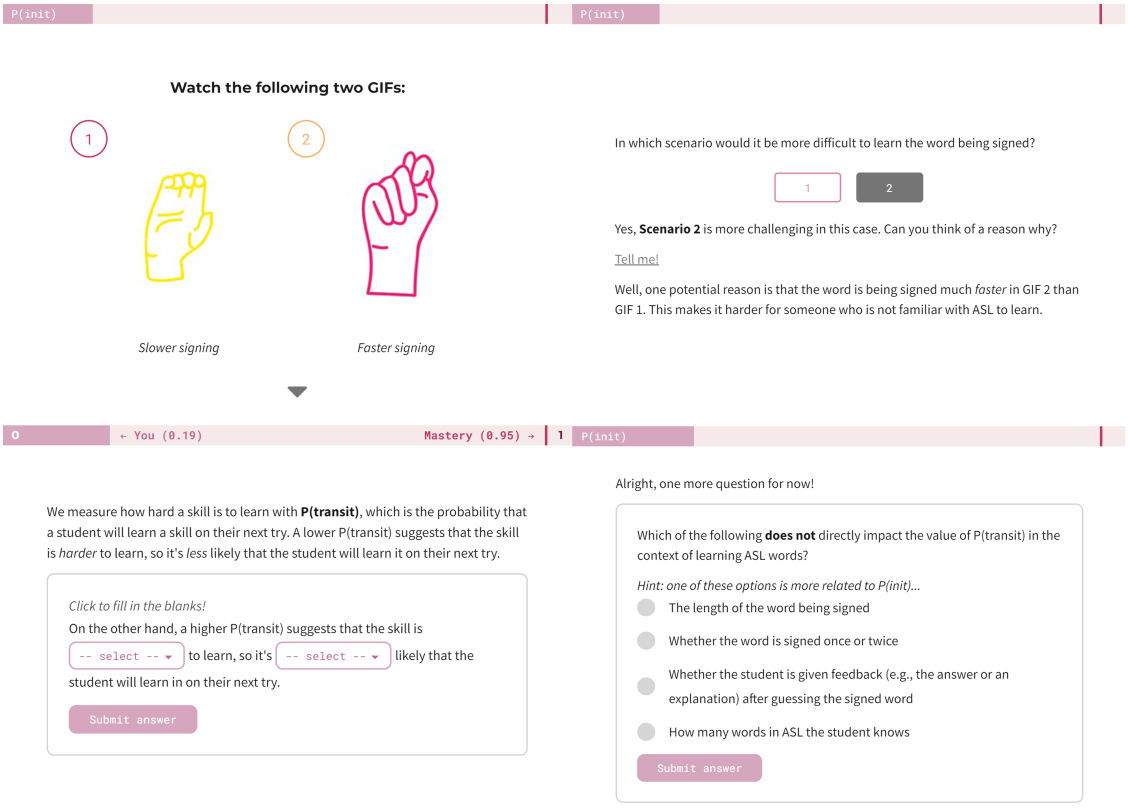
Hence, we did not end up using cognitive walkthroughs during the design process, but it seemed the combination of user studies and heuristic evaluations already provided a sufficiently robust scheme for user testing my BKT explainable prototype.

4.2 Explainable Implementation

To implement my explainable, I created a web application coded entirely in JavaScript/jQuery (version 3.5.1), HTML, and CSS. Throughout the implementation process, we continued to iteratively test and revise our design with participants until we reached a point of diminishing returns in which no new major functionality issues arose. Then, I published my finished BKT explanation, ultimately concluding the design process (Figure 4.1). The final design is a dynamic, interactive, publicly accessible explainable: https://catherinesyeh.github.io/bkt-asl/.

As previously mentioned, my explainable consisted of three main sections: 1) introducing each BKT parameter, 2) showing BKT in action through an ASL mini game, and 3) prompting deeper exploration of the algorithm with limitation modules. At the end of the explainable, we also provide a list of resources in case participants are interested in learning more about ASL.

Figure 4.7: Final P(transit) Module

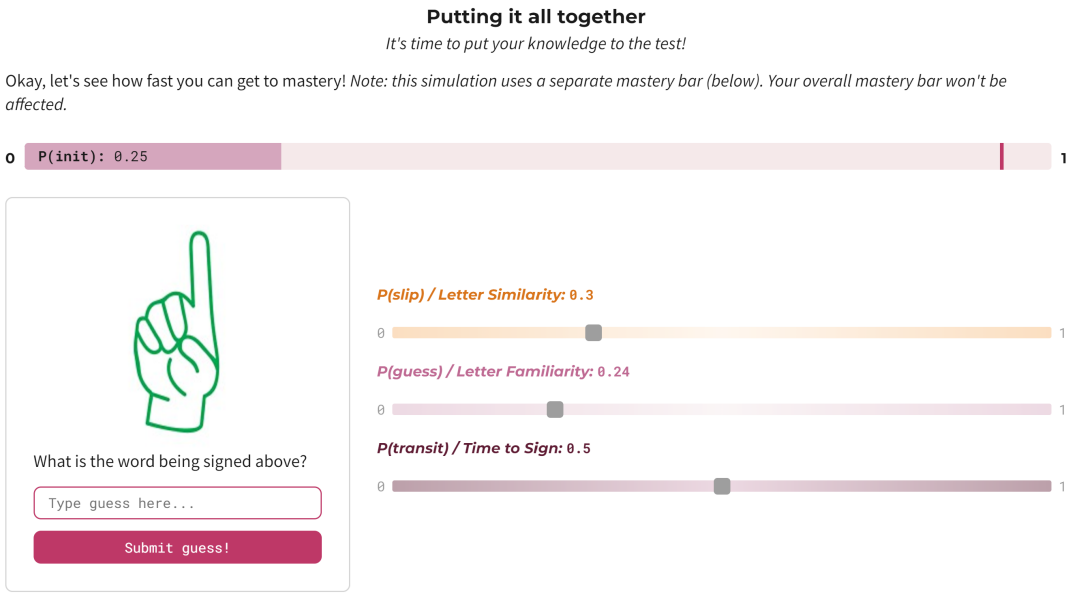


4.2.1 Parameter Introductions

One sample parameter module is shown in Figure 4.7, which introduces users to **P(transit)** and corresponds to the earlier prototype screens included in Figure 4.3. In this module, participants first see two GIFs of a word being fingerspelled at different speeds. By completing the activities below, they see how the speed of signing can make it more/less difficult for students to learn a signed word. Then, users can connect this idea to **P(transit)**, which expresses the probability of a student learning a skill on their next try (Table 2.2).

As illustrated by the pink bars at the top of each screen (Figure 4.7), we ultimately decided to make the main mastery bar horizontal instead of vertical as previously prototyped (Figures 4.3, 4.4). This allowed us to maximize the presentable content on each screen without overwhelming users and created a more natural layout considering the paginated organization of our final explainable. The main mastery bar will gradually increment toward the desired threshold of 0.95 as participants progress through the explainable. Similar parameter modules with interactive activities and questions were constructed for **P(init)**, **P(guess)**, and **P(slip)** as well.

Figure 4.8: Final BKT Simulation Activity



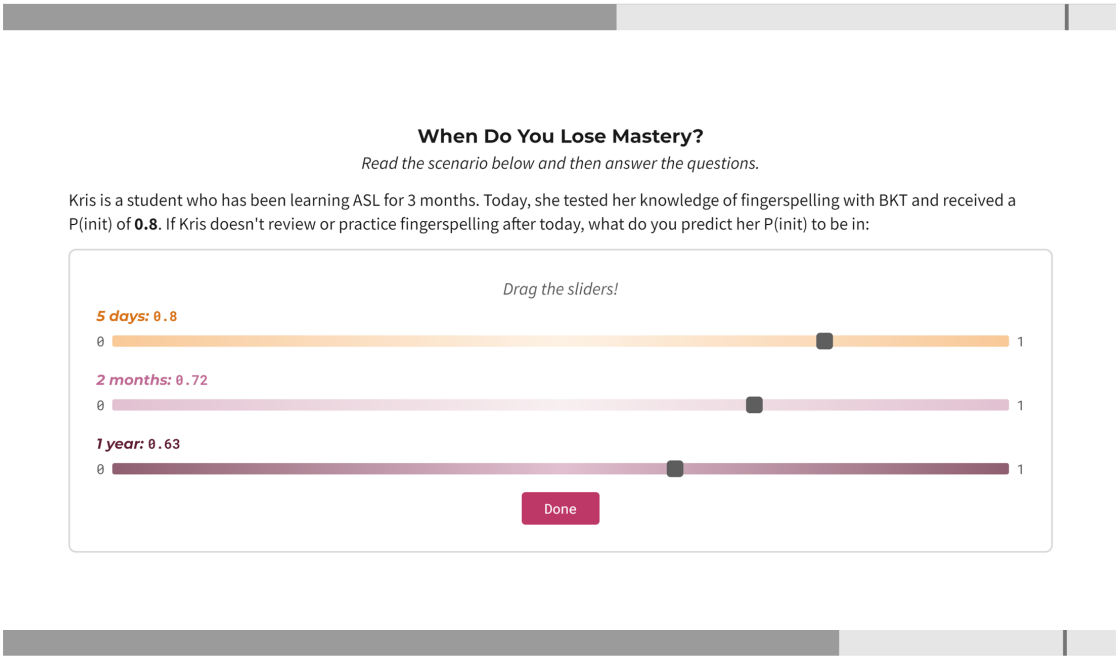
4.2.2 BKT in Action

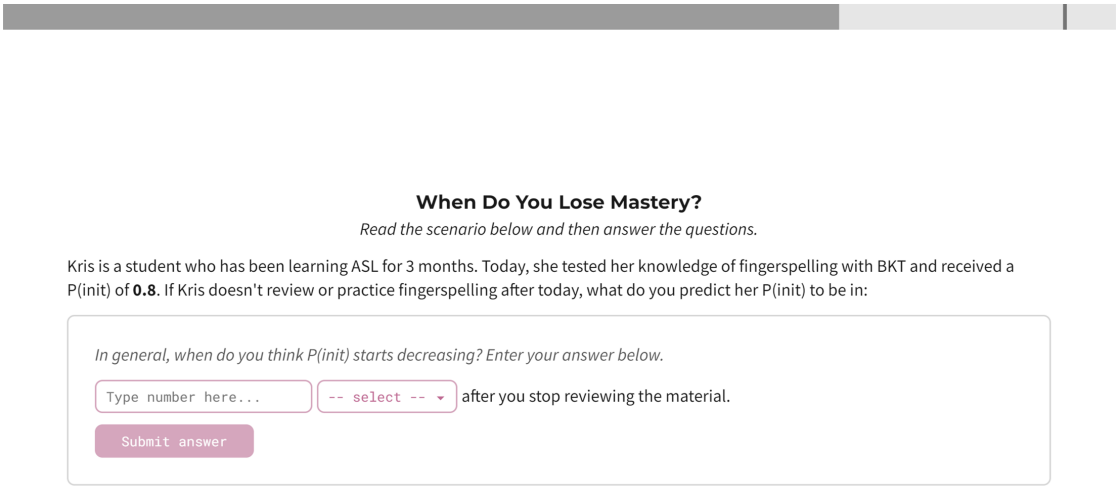
Next, we asked participants to put their knowledge to the test through a mini BKT simulation, which brings together all four BKT parameters. First, participants are asked to complete a parameter matching activity (Figure 4.5) to confirm their understanding of the different BKT parameters in the context of ASL.

We map **P(slip)** to *letter similarity* in ASL because when students see letters that are more similar, they are more likely to "slip up" and make a mistake. Similarly, we map **P(guess)** to *letter familiarity* in ASL because when students see letters that are more familiar to them, they are more likely to guess signed words correctly. Finally, we map **P(transit)** to *time to sign* in ASL because if words take longer to sign (i.e., are signed slower), they might be easier for students to learn. **P(init)** is not available as an option in this matching activity to reflect the fact that it is fundamentally different from the other three BKT parameters and is the only probability whose value is updated as students complete their learning exercises.

Then, participants can put their ASL skills to the test by guessing different fingerspelled words until mastery is achieved (Figure 4.8). Users are encouraged to play the game multiple times to explore how different parameter settings, adjusted via the provided sliders, affect the ease of attaining mastery. While the simulation is active, a separate mastery bar is used to illustrate how **P(init)** changes with each answer. The main mastery bar at the top of the screen (see Figure 4.7) is grayed out to indicate that participants' overall progress will not be affected. An earlier version of this activity can be seen on the left side of Figure 4.4.

Figure 4.9: Final Forgetting Limitation Module





4.2.3 Limitation Modules

Finally, after completing our ASL guessing game, we ask participants to complete four modules targeting various limitations of BKT (Figure 4.4), as described in Section 4.1.2. For example, in the "When Do You Lose Mastery?" module illustrated in Figure 4.9 (see Figure 4.4 for earlier prototype), participants are asked to assess the magnitude of **P(init)** at different points in time. The goal of this module is to demonstrate how BKT does not account for forgetting, which may bias its estimates of mastery. Again, the main mastery bar is grayed out on these screens to distinguish participants' overall progress from each module's activities (Figure 4.9).

Another limitation module, "What Causes Unexpected Model Behavior?", prompted users to explore BKT parameter values that result in model degeneracy (Figure 4.6) [33]. By completing this module, participants learn that if **P(guess)** and **P(slip)** are set to a value ≥ 0.5, BKT might predict a greater increase in student mastery for incorrect answers rather than correct answers. Thus, the stricter bounds of 0.3 for **P(guess)** and 0.1 for **P(slip)** are typically enforced to prevent degenerate behavior in BKT systems [27].

The "How do Incorrect Answers Impact Mastery?" module demonstrates how even incorrect answers typically yield an increase in mastery; you can walk through the math yourself by plugging different parameter values into Equations 2.1-2.4. This is not necessarily a flaw of BKT, but an interesting characteristic that reflects how in real life, every answer, whether wrong or right, can be viewed as a learning opportunity that brings you one step closer to mastering a skill.

Our final module, "What is the Role of Speed in BKT?", illustrates how BKT does not account for time when estimating student mastery [45]. That is, if two students get the same score on a test, but one finishes faster than the other, BKT would assign them the same **P(init)**, overlooking the possibility of speed as an indicator of recall/fluency and thus mastery.

4.3 Explainable Evaluation

After completing the implementation phase, I assessed the effectiveness of our final BKT explainable with a formal user study.

4.3.1 Pre- & Post-Tests

We used Qualtrics to design pre- and post-tests to accompany our BKT explainable in a remote format. For the full list of pre-/post-test questions used in this study, see Appendix B.

Because we did not expect most participants to come in with prior knowledge about BKT, as my explainable is targeting non-expert users, our pre-test did not serve as a baseline for participant post-test scores. Thus, this study did not involve conducting a controlled experiment and comparing results across different experimental groups. Instead, our **pre-test** consisted mainly of self-report questions capturing participant demographics (e.g., age, gender, education level, etc.) and math and computer science (CS) background. Prior work suggests that educational level impacts how users learn from post-hoc explainables [114], so we decided to include items to assess our participants' educational background and confidence. Math/CS background questions were adapted from [42, 55]

and based on Bandura's guide for constructing self-efficacy scales [12]. Self-efficacy measures self-reported beliefs that the participant *can* accomplish a task, not necessarily that they will accomplish it. Pre-test questions were a mix of multiple choice, open response, and 5-point Likert scale ratings (from 1: strongly disagree to 5: strongly agree) [5]. We also asked participants about their familiarity with Bayesian statistics/BKT systems and general attitudes toward AI; these questions were based on [34, 113]. See Appendix B.1 for a list of all the pre-test questions used.

On the other hand, our **post-test** questions were inspired by [73], which outlines different evaluation methods for XAI systems. For example, to evaluate user *mental models* of BKT, the first section of our post-test consists of questions specifically targeting our BKT KCs (Section 3.2). See Appendix C.2 for the full mapping of these post-test questions to KCs. This portion of the post-test includes a mix of multiple choice, short answer, and longer free response questions to assess participant understanding of BKT and its potential biases/limitations. Many of our questions were similar to the vignette-style problems included in our CTA protocol (Section 3.1.1), mirroring the scenarios we present to participants throughout the explainable. If our explainable explains BKT effectively, participants should perform well on this section of the post-test.

After these more BKT-specific problems, we also included some Likert-scale questions [5] to assess the other evaluation categories suggested by [73]. For example, to measure *usability and user satisfaction* of our explainable, we adapted questions from the System Usability Scale [13] and similar scales such as [43, 51, 85], incorporating self-efficacy statements [12] where applicable. These questions encapsulated measures of perceived usefulness (PU) and perceived ease of use (PEOU), which are key elements of the Technology Acceptance Model (TAM), a theory from the information systems community for measuring acceptance of technological systems [64]. TAM is an especially popular method for measuring acceptance in e-learning research [98], so we thought it would extend nicely to our study of XAI for learning analytics systems. Additionally, PU and PEOU have been shown to be strong drivers of user adoption of technology for practice [28]. Some applications of TAM also capture intention to use [64], but this was more difficult to achieve in my study considering the lab setting and the fact that our student participants were not active users of BKT systems.

Instead, we ventured beyond TAM with the remainder of our post-test, measuring *user trust* of BKT (based on our explainable) with a modified version of the six-construct scale from [18], which assesses system competence, integrity, benevolence, transparency, and more. Finally, we compared *user attitudes* toward AI algorithms more generally before and after completing our explainable using the same set of questions from our pre-test [34, 113]. Participants were given space to provide any other study feedback as well. See Appendix B.2 for a list of all the post-test questions used. Some questions are also discussed explicitly in Chapter 5.

4.3.2 User Studies

Once our pre- and post-test materials were complete, we ran a final round of Institutional Review Board (IRB) approved **user studies** with nine Williams College students to test the efficacy of our explainable. Students were recruited using the all-campus daily message system, representing all class years and a wide array of majors. Participants were asked to fill out a consent form, complete

the pre-test survey, step through the explainable, and finally answer the post-test questions, all via Qualtrics. In order to assess the explainable's effectiveness, we asked participants to complete the post-test from memory, without relying on the BKT explainable or other external resources. User studies lasted an average of 40 minutes in duration and participants were compensated with $20 Amazon gift cards for their time. The studies were completed remotely via Zoom, where I would first introduce the task and provide general instructions before then remaining in the video chat for the duration of the study to answer any questions from participants.

4.4 Summary

This chapter describes the methods used throughout my thesis to design and implement our post-hoc explainable for Bayesian Knowledge Tracing. I first discuss the brainstorming, prototyping, and user testing process used to create our initial explainable design. During the brainstorming stage, we decided to based our explanation of BKT around American Sign Language to demonstrate the real-world applicability of the algorithm and allow users to learn a new, useful skill while competing my BKT explainable. Collecting feedback on my design via user studies helped to improve its overall functionality and user experience.

After the final hi-fi prototype was complete, I ultimately implemented our BKT explainable as an interactive web application using JavaScript, HTML, and CSS. The explanation begins by introducing each of four BKT parameters separately, using various interactive activities and self-explanation prompts to encourage active learning. We then bring the parameters together with a mini BKT simulation activity to show the algorithm in action on a smaller scale. Finally, we have a series of modules to teach some of BKT's key limitations and flaws, including model degeneracy and the fact that BKT does not account for the possibility of forgetting.

At the end of the chapter, I explain our process for developing pre- and post-tests, which will be used with our explainable to measure user learning and other behavioral outcomes. The pre-test consists of a brief demographic survey and some questions targeting our participants' math and CS backgrounds. In our post-test, we first ask questions to assess the BKT mental models gained by participants from completing our explainable; these questions specifically target the KCs identified previously with CTA. Then, we collect additional measures including user satisfaction, trust toward BKT, and attitudes toward AI more broadly. Lastly, we incorporated all these materials into a Qualtrics survey and ran user studies to assess the overall effectiveness of our novel XAI framework.

Chapter 5

Results & Discussion

For my thesis, I applied approaches from Backward Design, user-centered design, and best practices from teaching & learning theory to create an evidence-based post-hoc AI explainable, with BKT as the example AI algorithm. Our results evidence that the first step toward constructing post-hoc XAI should be determining the basic building blocks of understanding experts of that AI algorithm already possess, before moving to the next step of designing assessments and explanatory activities targeting each of the resultant knowledge components.

In this chapter, I discuss the insights from running a user study on my final BKT explainable through the lens of three research questions:

- **RQ1:** Does our explainable offer an effective explanation of BKT?
- **RQ2:** What factors impact successful learning with our explainable?
- **RQ3:** How does our explainable impact user attitudes toward BKT and AI more broadly?

5.1 Does Our Explainable Offer an Effective Explanation of BKT?

By analyzing user performance on our post-test, we determined that participants were able to gain a solid understanding of BKT through completing this study. None of the nine students we recruited had prior experience with BKT, but they ultimately attained an average overall post-test score of 87% as a group, shown in Figure 5.1, attesting to the effectiveness of our explainable.

5.1.1 Explainable Effectiveness by Knowledge Area

Although our explainable was effective overall, it is clear that certain topics were harder for participants to grasp than others, as illustrated by the average scores with standard deviation error bars in Figure 5.1. Participants struggled the most with the **Identifying Priors** knowledge area, scoring an average accuracy of 72.2%. In particular, participants found it challenging to recall the names

Figure 5.1: Participant Post-Test Scores by Knowledge Area

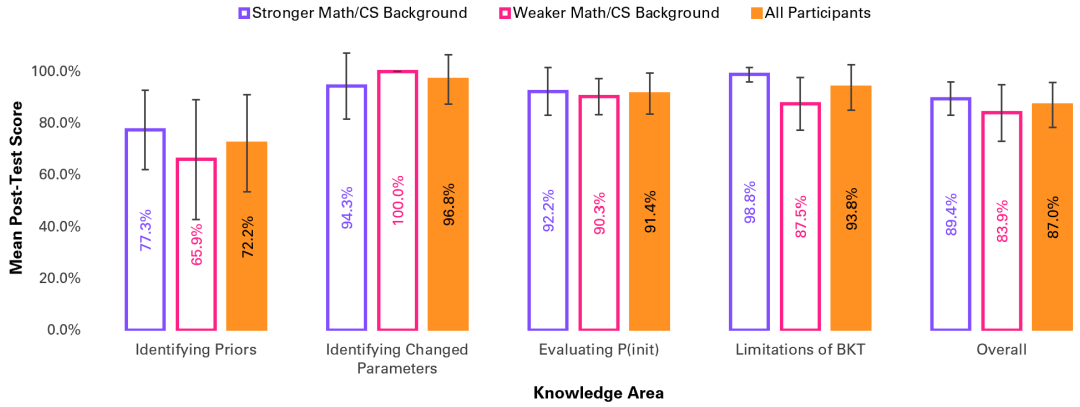


Table 5.1: Post-Test Problem Probing Model Degeneracy

Pinto is taking a strange multiple-choice exam where **P(guess)** is 0.67. They also drank too much coffee and find it difficult to fill in the bubbles correctly on the answer sheet (i.e. They fill in "C" when they want to select "A"), so their **P(slip)** is 0.5.

1. Consider the given values of **P(guess)** and **P(slip)** and explain how this may contribute to unexpected model behavior.

and definitions of the four BKT parameters and what the acceptable bounds were for **P(slip)** and **P(guess)**. We had previously thought that **Identifying Priors** would be one of the simpler and more intuitive knowledge areas of BKT to learn, but our results suggest otherwise, evidencing that perhaps more attention should be paid to helping system users solidify these foundational concepts before moving on to more complex skills. It may also be the case that for probabilistic algorithms like BKT, predicting system behavior is more subjective and challenging than a binary classifier for instance, so perhaps this need not be a main goal of such post-hoc AI explanations.

On the other hand, participants performed the best on the **Identifying Changed Parameters** knowledge area, scoring an average accuracy of 96.8%, which suggests that recognizing changes in parameter values may be a more intuitive skill for BKT non-experts. Continuing the discussion about probabilistic algorithms from above, it may be more important for everyday users to recognize trends in the system's behavior (e.g., increase or decrease) rather than predicting BKT's precise output (e.g., **P(init)** = 0.78) anyway.

Participants also scored highly on questions pertaining to **Evaluating P(init)** and **Limitations of BKT**, attaining average accuracies of 91.8% and 93.8%, respectively. We originally presumed these two knowledge areas would be more difficult for users, but it seems that the extra time and activities dedicated to them in the explainable helped participants successfully learn these concepts. Our participants may have also had stronger intuition for these knowledge areas than we anticipated,

Table 5.2: Participant Post-Test Ratings of Explainable Design (1: Strongly Disagree - 5: Strongly Agree)

Explainable Design	**Mean Rating (SD)**
I thought the explainable was easy to follow.	4.78 (0.44)
I found the various sections in this explainable were well integrated.	4.89 (0.33)
I found this explainable engaging and interesting.	4.89 (0.33)
I enjoyed completing this explainable.	4.67 (0.50)
Average:	**4.81 (0.11)**

which seems promising toward the goal of revealing the nuances and potential limitations of complex algorithms through the means of XAI. Questions in the **Evaluating P(init)** and **Limitations of BKT** areas did tend to result in less uniform answers, but it should be noted that these problems are inherently more open-ended by design.

In the **Evaluating P(init)** knowledge area, the most challenging questions revolved around deciding skill mastery thresholds for different situations involving BKT. For example, we asked participants to assign appropriate thresholds for assessing mastery for a math worksheet versus a medical school entrance exam. Many participants were hesitant to stray from the typical threshold of 0.95 [27], even though it was mentioned that this value could vary across different scenarios. It is possible that adding more examples in our explainable to illustrate the variance of mastery thresholds would have helped participants feel more comfortable with choosing alternative cutoff values, and this is something to consider when designing future post-hoc AI explanations.

In the **Limitations of BKT** knowledge area, our participants often surprised us with their ability to apply knowledge from the explainable to uncover new biases and limitations of the algorithm. For instance, one of our post-test questions, shown in Table 5.1, was targeting a simple explanation about model degeneracy [33] due to abnormally high values of **P(guess)** and **P(slip)**. However, a few participants took this one step further, detailing how "the model doesn't take in the person's state when taking an exam, it will assign the same score to the highly caffeinated pinto as to the normal pinto" or how "P(guess) and P(slip) are statistics representative of the subject at a stable mindset but the external caffeine factor has altered the conditions, causing potentially unexpected behavior and results." Considering that we did not mention anything explicitly about BKT's relationship to students' mental states in our explainable, it is encouraging and fascinating that participants made this connection on their own.

5.1.2 Explainable Effectiveness by Participant Ratings

Participant post-test ratings of our explainable were also largely positive, as illustrated in Tables 5.2 and 5.3. For this section, we dropped one of our participants, as they likely inverted their Likert scale ratings for strongly disagree and agree (e.g., their responses were recorded in Qualtrics as all "strongly disagree", while all other participants marked "strongly agree" or "somewhat agree").

Table 5.3: Participant Post-Test Ratings of Explainable Effectiveness (1: Strongly Disagree - 5: Strongly Agree)

Explainable Effectiveness	**Mean Rating (SD)**
I thought this explainable provided an effective explanation of BKT.	4.67 (0.50)
I could generally explain how BKT works to another person.	3.67 (1.12)
I feel comfortable using BKT systems after completing this explainable.	3.89 (0.60)
I am interested in learning more about ASL.	4.22 (1.09)
Average:	**4.11 (0.44)**

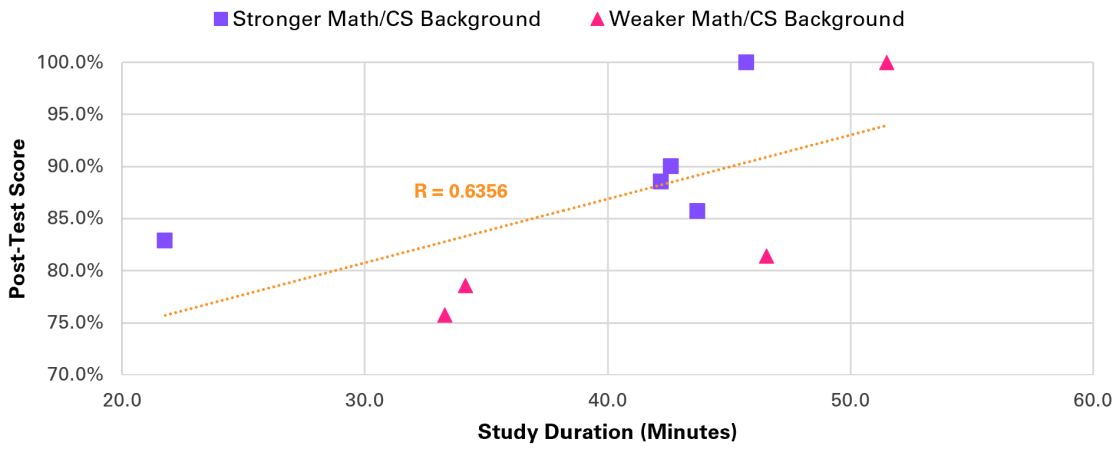
The *explainable design* categories accrued a mean rating of 4.81 (Table 5.2), and several participants mentioned their satisfaction with the explainable when given the opportunity to provide open-ended feedback. For example, one participant reported, "So aesthetic and easy to follow! This was really fun." Another wrote, "I really liked the dynamic design of the explainable, and it was very interactive. It was easy to process and digest even with minimal background in computer science or in ASL." Other participants offered helpful suggestions for further improving the explainable: "I wish that there was a little glossary at the top that you could click on to remind you what the different P values meant, because those definitions only showed up once. It's helpful for me to have optional reminders." This suggestion in particular speaks to the challenge of balancing cognitive loads [87] and scaffolding theory [52]—offering ample support to students as they learn a new skill—when designing AI explainables.

Questions in the *explainable effectiveness* categories accrued a slightly lower mean rating of 4.11 (Table 5.3), but notably, our participants strongly agreed with the statement "I thought this explainable provided an effective explanation of BKT," which received a mean rating of 4.67. It is also encouraging that our participants expressed interest in learning more about ASL after completing our explainable, which yielded a mean agreement rating of 4.22. Since this was all of our participants' first exposure to BKT, it is understandable that they might not have felt 100% confident about explaining or using the algorithm after completing the explainable alone. One participant also mentioned the desire to know "more about how [BKT] is used in the real world, because that's really interesting... For example: how does BKT come up with those numbers?" This implies that there is still more work to be done in terms of teaching real-world applications of AI algorithms. However, as demonstrated by the overall high post-test ratings and scores obtained by participants, our explainable was successful in providing non-experts with a strong basic understanding of BKT.

5.2 What Factors Impact Successful Learning With Our Explainable?

As prior research on post-hoc explainables suggests that user backgrounds can influence algorithmic understanding [114], we originally planned to investigate trends within participant subgroups. While

Figure 5.2: Participant Post-Test Scores vs. Study Duration



we collected data on demographic groups such as age and education level (as discussed in Section 4.3.1), our subject pool ended up consisting entirely of undergraduate students at Williams College, so this information did not prove to be significantly varied. Instead, we chose to investigate time-on-task and math/CS background; how we chose to define the latter is discussed below.

Research from the learning sciences suggests that increasing student **time-on-task** should lead to increased learning, with some qualifiers [97] proposing study duration (i.e., the amount of time participants took to complete the study) as a relevant factor. Because the study was self-paced, there was inevitably some variability in completion times. As shown in Figure 5.2, which plots each participant as an individual data point, we discovered a relatively strong positive correlation between study duration and post-test scores (R = 0.6356). This result demonstrates that participants who spend more time engaging with the explainable are more likely to successfully learn the material, corroborating the findings from [97].

Additionally, I decided to compare performance between participants who had stronger and weaker **math/CS backgrounds**. These groups were formed by taking into account both self-reported confidence ratings with math/CS and participant pre-test scores. From Figure 5.2, we can see that students with weaker math/CS backgrounds did take longer to complete the study on average, but they did not necessarily perform worse than their peers; in fact, one of the two students who achieved 100% on the post-test was in the weaker math/CS group. Thus, it seems that anyone, regardless of math/CS background, can attain comprehensive understanding of AI algorithms like BKT, as long as they commit the effort. In other words, prior math/CS proficiency is not a prerequisite to successful learning with our explainable.

A similar trend can be observed in Figure 5.1. Although participants with weaker math/CS backgrounds generally achieved lower average accuracies across our BKT knowledge areas, they did not perform noticeably worse than their peers and still achieved relatively high scores overall (weaker math/CS: 83.9% vs. stronger math/CS: 89.4% overall post-test accuracy). In fact, the weaker math/CS group performed better than the stronger math/CS group on questions from the

Table 5.4: Participant Post-Test Ratings of BKT Trust (1: Strongly Disagree - 5: Strongly Agree)

User Trust of BKT	**Mean Rating (SD)**
BKT provides accurate estimates of skill mastery.	3.56 (0.73)
BKT provides reliable, unbiased estimates of skill mastery.	3.56 (0.73)
BKT systems are helpful to students/teachers and reflect their best interests.	4.00 (0.50)
BKT's estimates of skill mastery are logical and easy to understand.	3.89 (0.93)
I would use BKT to assess skill mastery in different settings.	3.33 (1.00)
In general, I trust BKT's estimates of skill mastery.	3.78 (0.44)
Average:	**3.69 (0.25)**

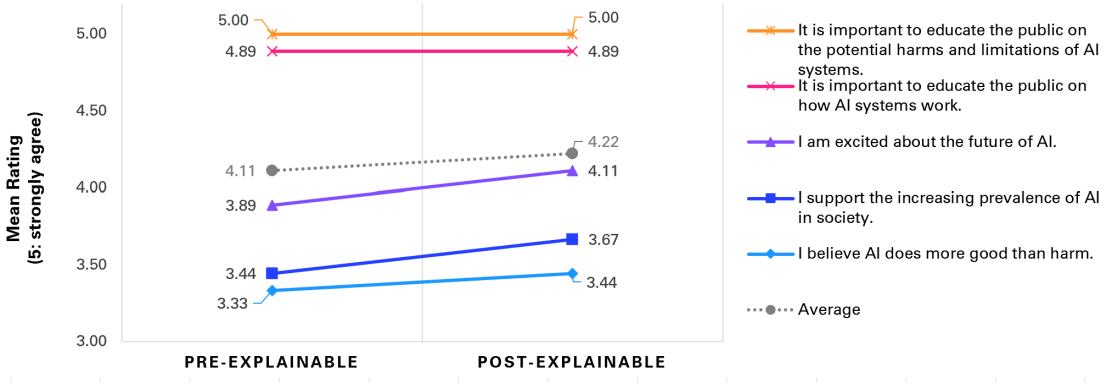
Identifying Changed Parameters knowledge area, with all participants achieving an impressive 100% accuracy. These results evidence how our explainable is similarly effective to all non-expert users, regardless of math/CS background, achieving our goal of designing a post-hoc AI explanation that is accessible to everyone.

One small caveat is that these observations could also potentially be explained by confounding variables. For example, math/CS background may be a confounder of the positive correlation between study duration and participant post-test scores, as this variable could impact both time-on-task and learning. Additionally, due to the relatively subjective definition used to split participants into groups based on stronger/weaker math/CS backgrounds, we may have not accurately captured the true math/CS proficiency of our participants. There could also be unmeasured confounders here, such as math/CS inclination, as our self-report questions related mostly to confidence and knowledge rather than interest in these subjects.

5.3 How Does Our Explainable Impact User Attitudes Toward BKT and AI More Broadly?

Finally, I looked at how my explainable impacted user attitudes toward BKT and AI more broadly. Table 5.4 shows participant post-test ratings targeting *user trust of BKT*, which averaged to 3.69 across the board. This suggests that participants do generally trust BKT systems, but they would not place blind faith in the system and are a bit weary of its behavior in certain scenarios. In particular, the statement which received the lowest average rating of 3.33 was "I would use BKT to assess skill mastery in different settings," indicating an understanding that BKT's behavior is not equally reliable in all situations. On the other hand, the statement which received the highest average rating of 4.00 was "BKT systems are helpful to students/teachers and reflect their best interests," demonstrating an overall positive outlook toward educational technologies and augmenting learning environments with AI. This tension between wanting to trust BKT and hesitating because of its

Figure 5.3: Changes in Participant Attitudes Toward AI



various limitations and biases was also reflected through participant feedback. One participant wrote, "From my understanding of BKT, it seems like many of the factors that impact learning (like forgetting material over time, time needed to complete a task, or learning differences) are not accounted for. It still seems cool though!" Another participant reported how "I feel like there are external factors which may not always be accounted for, and it feels weird that P(init) doesn't change even though I may forget the skill/fact over a course of time." But because one of our objectives was to encourage users to grapple with BKT's algorithmic flaws and question its predictions when appropriate, these results are another indicator of our explainable's success.

In terms of *user attitudes toward AI*, I noticed a slight increase in participant ratings throughout the study, as the pre-explainable mean rating for these questions was 4.11, compared with the post-explainable mean rating of 4.22 (Table 5.3). Notably, before completing the explainable, all of our participants already strongly agreed with the statement, "It is important to educate the public on the potential harms and limitations of AI systems," and this remained true post-explainable as well. The second most highly agreed with statement with an average rating of 4.89 (both pre- and post-explainable) was "It is important to educate the public on how AI systems work." Interestingly, this statement was rated slightly lower than the former on average, suggesting that perhaps it would be more beneficial for XAI systems to target algorithmic flaws and biases over inner workings of ML systems when the audience is the general public. After all, it is logical that for everyday users, knowing when AI systems work well versus when they might fail would be more helpful that knowing precisely how the algorithm itself works.

Additionally, the fact that the average participant ratings for the last three statements in Table 5.4 all increased over the course of the study is indicative of a generally optimistic view toward the future of AI. Despite learning about the potential limitations of these systems, our participants still adopted largely positive attitudes toward AI, demonstrating that increasing algorithmic understanding will not necessarily also decrease user trust of ML systems, as suggested by [7]. Of course, some of our participants still had their concerns about the increasing prevalence of AI in society: "I am concerned by how AI systems can easily become biased, because the data they are trained on is also biased." However, most students remained hopeful about AI's potential for societal impact: "I

think AI has a lot of benefits that will allow greater efficiency for certain processes previously done by humans. However, I am concerned about whether unemployment would increase or there would be other jobs to substitute the existing, soon to be outdated ones. Nevertheless, I think it is a very impressive system that humans have created and I am excited to see how it is applied and unfolds and will become more integrated into our lives." In general, I was impressed by the nuanced understanding of BKT and AI participants demonstrated in this study, and these open-form responses only strengthened the observations made in Section 5.1.1 about participant post-test performance.

5.4 Limitations

Limitations of this work largely arise from limitations of the methods. Our CTA protocol 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 during this protocol. For example, there is always the possibility that our KCs may not be completely accurate due to experts having processed information or completed certain steps to a problem in their head (e.g., recalling the definition of a parameter or synthesizing the information from a vignette), but again, this is a potential limitation of any think aloud protocol. Ideas for more tangible, observable ways of capturing KCs are discussed in Section 7.2.

Our sample sizes were also limited throughout this study, but the consistency across our participants increases confidence in our findings. For the CTA portion of this work in particular (Section 3.1), our small sample size stems from the currently sparse supply of student BKT experts. Furthermore, for usability testing, the optimal sample size is actually four or five participants, as observing more users typically does not uncover additional significant usability problems, ultimately leading to diminishing returns [100]. However, all of the participants recruited for my thesis were undergraduate students from the same school, Williams College, creating a narrow subject pool, so our results may not be generalizable to outside the population studied. As previously discussed in Section 4.3.1, we also did not run a controlled experiment to evaluate the explainable, so our results discuss trends and correlations rather than causal relationships. The possibility of confounding variables (e.g., time-on-task and math/CS background) should be taken into account as well when interpreting and building off this work.

Additionally, while these KCs apply to BKT, it may be difficult to generalize them directly to another algorithm, although the CTA method itself does extend to other contexts. One challenge that may be encountered with other algorithms is that it is difficult to produce reliable KCs for more subjective questions where experts tend to diverge in thought process, such as those involving the limitations of BKT. However, it should be noted that CTA is traditionally not used to probe the cognitive processes of such open-ended questions and is more typically used in structured scenarios like performing statistical analyses [68]. Nonetheless, a similar think aloud protocol with subject matter experts could be applied to determine the KCs of other AI algorithms. Similar knowledge areas could also be targeted, such as identifying initial/changed values of parameters or the limitations of an algorithm, if applicable. In this work, we do not claim that our knowledge components are necessarily generalizable to other algorithms, but that the methods to find these KCs are.

Another limitation is related to the critique of post-hoc explainables more generally. As the quantity and variety of algorithms in our daily lives continues to increase, how do we determine which algorithms need a post-hoc explanation? Providing a post-hoc explanation for every complex algorithm is not a scalable solution, so it is important to consider the optimal approach to this problem. Similarly, even once we have a complete post-hoc explainable, there remains the challenge of how to integrate them effectively into real-world settings. XAI systems are particularly critical for high-stake domains like healthcare and automated transportation [44, 61], but it is often difficult to deploy them in industry. There is little understanding of how companies and organizations actually use these post-hoc explanations, and frequently, they are only made available to internal stakeholders like ML engineers (i.e., the people building and debugging AI systems) rather than end users [19]. Thus, there is still work to be done in terms of addressing this gap between the idealized goals of post-hoc AI explainables and how they are currently used in practice.

5.5 Summary

To summarize, the results from my user studies evidence that our novel, learning sciences inspired XAI methods framework did generate an effective, accessible post-hoc explanation for BKT. Participants generally enjoyed completing my explainable and thought the design was effectual for learning. Concepts in certain knowledge areas were harder for our non-expert participants to grasp, but they were also able to successfully learn and apply more complex, nuanced pieces of BKT. In addition, we determined that effort, rather than prior expertise, is key to effective learning with our explainable, as even participants with weaker math/CS backgrounds were able to attain high scores on our BKT post-test, sometimes outperforming their peers. Finally, my results reveal more skepticism toward blindly trusting and using BKT, indicating a deeper understanding of the algorithm's potential flaws and biases, but overall, our participants reported fairly optimistic attitudes toward the growing number of AI applications in society. These results should be interpreted with caution, however, due to the possibility of confounding and other limitations mentioned above.

Chapter 6

Additional Study: The Effect of Explanation on User Outcomes

This chapter describes an additional study we conducted to explore the effect of algorithmic transparency on user outcomes more deeply, expanding on our final research question:

- How does our explainable impact user attitudes toward BKT and AI more broadly?

Specifically, we wanted to know how user understanding of the algorithm's limitations impacts their attitudes toward BKT and AI. To examine this question, I varied the amount of information presented about BKT's limitations to users in my explainable and then investigated how this impacted their perceptions of decisions made by AI algorithms vs. human decision-makers.

6.1 Motivation & Prior Work

From the results of our first study (as discussed in Chapter 5), we determined that our learning-sciences inspired approach for creating post-hoc AI explainables was effective in imparting key concepts about BKT to non-expert student users. In particular, our participants demonstrated a sophisticated understanding of the algorithm's potential flaws and limitations. Consequently, participants were hesitant to place unquestioned trust in BKT (Table 5.4), but still expressed optimism about the increasing prevalence of AI in society (Figure 5.3). Inspired by these key observations, we were interested in investigating further how algorithmic understanding impacts user outcomes (e.g., perceptions of trust and fairness).

To do this, we combined the idea of varying the level of transparency in post-hoc AI explanations [58] and measuring user outcomes via decision-making scenarios [63]. Some research suggests that increasing algorithmic transparency can lead to more realistic trust in ML models [7]. However, prior work by [58] demonstrates that greater transparency does not guarantee positive user outcomes and may lead to decreased trust and satisfaction. Similarly, other studies such as [22, 89] have found that more transparent models can lead to increased overreliance on AI systems. Thus, we were interested

to explore the impact of varying algorithmic transparency in our explainable on user behavior and decision-making.

Taking inspiration from [69], we wanted to construct and compare different BKT explanations for this study. But rather than varying the form of the explainable (e.g., visual vs. verbal vs. decision tree) [69] or the interactivity level as in [114], we decided to adjust the content distribution. Specifically, we aimed to investigate whether varying the amount of information provided about BKT's limitations would affect user perceptions of algorithmic vs. human decision-makers.

We chose to vary algorithmic transparency through a limitations lens because the existing XAI literature largely focuses on the effect of elucidating the inner workings of AI models (i.e., how the system itself works) [14, 22, 58, 61, 69, 89]. However, we believe that it is also crucial for XAI systems to address the potential shortcomings and biases of AI algorithms, as this will allow for more reliable decisions to be made with the assistance of AI [35]. The design of this experiment allows us to explore the veracity of this claim and justify it empirically. Additionally, educating end users about algorithmic limitations can help mitigate the dire consequences that may arise when using AI-mediated decision-making systems in high-stakes fields such as healthcare and criminal justice [61].

The observation from our first study that participants already seemed highly attuned to the limitations of BKT (Figure 5.1) also made us curious whether reducing the amount of information about the algorithm's flaws would impact their post-test performance and other behavioral outcomes such as perceived fairness and trust. In [63], the researchers measured perceptions of algorithms by presenting users with different scenarios involving managerial decisions. The decision-maker (either algorithmic or human) was manipulated in each scenario before measuring users' perceived fairness, trust, and emotional response. Ultimately, this study showed that on "human" tasks like hiring and work evaluation, algorithmic decisions were perceived as less fair and trustworthy than human decisions, while also evoking more negative emotion from participants [63].

For my thesis, we were curious how the algorithmic understanding conveyed by our BKT explainable would impact user perceptions of similar algorithmic vs. human decisions. Hence, we developed this study to answer the following set of new research questions:

- **RQ1:** Are the results from our first study generalizable to other college students?
- **RQ2:** How does decreasing the transparency of BKT's limitations affect algorithmic understanding?
- **RQ3:** How does decreasing the transparency of BKT's limitations affect perceptions of algorithmic fairness and trust?

The methods described in the following section can be extended to other algorithms and contexts, but to build off and take advantage of my prior work, we again use BKT as our algorithm of interest.

6.2 Methodology

To conduct this study, I began by creating alternate versions of my explainable with varying degrees of information about BKT's limitations. Then, I designed different scenarios with human or AI decision-makers to measure user perceptions of fairness and trust. Finally, I updated the Qualtrics survey from our original study (Section 4.3) to incorporate these additional materials and ran a similar user study to test our research hypotheses.

6.2.1 Explainable Conditions

We designed three explainable conditions for this experiment: (1) Long Limitations, (2) Short Limitations, and (3) No Limitations. In all conditions, the limitation section followed the culminating BKT simulation activity (Section 4.2.2) in our explainable. Participants were randomly assigned to an explainable condition.

Long Limitations

In the *Long Limitations* condition, we showed participants our original BKT explainable (available here: https://catherinesyeh.github.io/bkt-asl/), which includes the four, detailed limitations modules as described in Section 4.2.3. Each module consists of various interactive activities and self-explanation prompts to demonstrate how BKT:

- does not account for forgetting when generating its estimates of skill mastery [32],
- may yield degenerate model behavior [33] if parameter values (i.e., **P(guess)** and **P(slip)** in particular) are out of range,
- increases its skill mastery estimates even if students answer a question incorrectly, and
- does not take time into consideration when estimating student skill mastery [45].

Short Limitations

In the *Short Limitations* condition, we showed participants a modified version of our BKT explainable with condensed limitation modules (available here: https://catherinesyeh.github.io/bkt-asl-v3/). For each of the four limitations described above, I summarized the information in a few paragraphs (compared with the multiple slides/pages in our original explainable) and replaced the more comprehensive activities with images or GIFs illustrating the same concepts. Figure 6.1 shows an excerpt of our shortened model degeneracy module; see Figure 4.6 for the full model degeneracy module in our original explainable.

In some ways, this experimental condition mirrors [114]'s work, as this version of our limitation modules is much more "static" than the original, which involved more engagement and active learning from the user. However, by only modifying the form of the limitation section in my explainable, we can investigate how reducing the interactivity and amount of information provided about BKT's

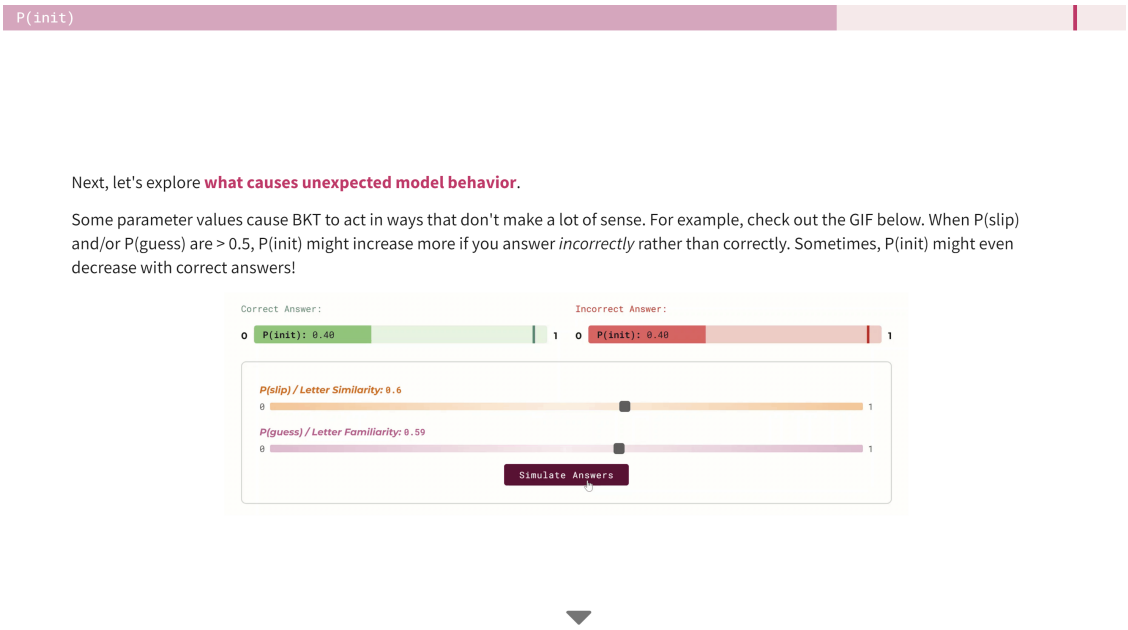


Figure 6.1: Shorter Model Degeneracy Module

limitations impacts user learning and behavioral outcomes. Specifically, this condition aimed to test whether placing less emphasis on algorithmic limitations would change the understanding and perceptions about BKT gained from our explainable.

No Limitations

In the *No Limitations* condition, we showed participants a modified version of our BKT explainable with no limitation modules (available here: https://catherinesyeh.github.io/bkt-asl-v2/). That is, after completing the mini BKT simulation (Section 4.2.2), participants finished the explainable and were immediately taken to the ending page.

This condition was meant to serve as a control, measuring the necessity and impact of providing explicit information about BKT's limitations in a post-hoc explanation of the algorithm. Because our participants were able to successfully extrapolate information about BKT's flaws and biases on their own in my first study (Figure 5.1), we were curious to see if this intuition would still arise after completing an explainable purely focused on the "inner workings" of BKT and how participant's trust and fairness ratings of decisions made by BKT would be affected.

6.2.2 Decision Scenarios

Next, we created four decision scenarios inspired by [63]. These scenarios were then incorporated in our post-test so we could see how our different BKT explainable versions impacted user perceptions of fairness and trust in algorithmic decisions (see Appendix B.2.2).

Table 6.1: Decision Scenarios Presented to Participants

Scenario Type	Scenario Context	Scenario
Low-stakes	First grade math worksheet **(AI)**	*General:* At an elementary school, first graders complete a math worksheet each week. An AI algorithm assesses their performance on these worksheets to decide whether they have mastered the material. *Specific:* Mars is a first grader at the elementary school. The AI algorithm evaluates their performance on this week's math worksheet.
	Second grade spelling worksheet **(Human)**	*General:* At an elementary school, second graders complete a spelling worksheet each week. Their teacher assesses their performance on these worksheets to decide whether they have mastered the material. *Specific:* Opal is a second grader at the elementary school. Their teacher evaluates their performance on this week's spelling worksheet.
High-stakes	Medical school entrance exam **(AI)**	*General:* 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. *Specific:* Clay applies to the medical school. The AI algorithm evaluates their performance on the entrance exam.
	Law school entrance exam **(Human)**	*General:* At a law school, first-year applicants must complete an entrance exam. To be recommended for admission, applicants must score highly on this exam. A human grader assesses their performance on the entrance exam. *Specific:* Emerson applies to the law school. The human grader evaluates their performance on the entrance exam.

However, instead of designing scenarios with tasks involving human or mechanical skills [63], we decided to design scenarios with *low-* or *high-stakes* decisions. All scenarios were intended to mirror real-world situations where BKT or other learning analytics algorithms might be used. Like [63], we also varied whether the decision-maker in each scenario was a *human* or an *AI algorithm*. Originally, we considered adding additional dimensions (e.g., school vs. non-school scenarios, decision vs. no decision presented, etc.), but due to limited resources and time, we ultimately opted for a simple 2x2 experimental design, paralleling [63]'s study.

I followed [63]'s approach of creating two-part scenarios written in the projective viewpoint [78], first providing general context about the situation before presenting a more specific instance of the scenario involving a fictional persona. The final scenarios we included in our post-test are displayed in Table 6.1. For each scenario type (i.e., low- or high-stakes), we aimed to create two comparable scenarios (e.g., first grade math worksheet and second grade spelling worksheet), which were then randomly assigned to a human or AI decision-maker. During our study, participants saw all four scenarios, but the order of presentation was randomized.

Perception Measures

To evaluate user perceptions of decision *fairness* and *trust* in each scenario, we adopted the questions used in [63]. Our full perception measures are available in Appendix B.2.2.

For example, to measure fairness in Scenario 1 (Table 6.1), we asked users, "How fair or unfair is it for Mars that the AI algorithm evaluates their performance on their math worksheets?" Similarly, to measure trust in this scenario, we asked users, "How much do you trust that the AI algorithm makes good-quality assessments of Mars' performance on their math worksheets?" Participants answered these questions by providing ratings on a 5-point Likert scale (e.g., "very unfair" to "very fair" or "extreme distrust" to "extreme trust") [5] and were also given space to optionally elaborate on their answers. These fairness and trust measures immediately followed their respective scenarios in our post-test.

The original study also included questions targeting emotional responses to decision scenarios [63]. However, we opted to omit these measures because this was less of a focus in our work and we did not want to further increase our study duration, which was already estimated to require one hour of our participants' time.

6.2.3 User Studies

After designing our new explainable conditions and post-test decision scenarios, I ran another round of IRB approved user studies to test our research hypotheses (Section 6.1). These user studies were very similar to those conducted in my previous thesis work, as participants were asked to fill out the pre-test survey, walk through our BKT explainable, and then answer the post-test questions (Section 4.3.2). Again, the whole study was coordinated and conducted via Qualtrics. Study duration varied by explainable condition (see discussion in Section 6.3.1), but was intended not to exceed 1 hour. All participants were compensated with $15 Amazon gift cards for their time.

Table 6.2: Participant Demographics

Demographic	**Participant Distribution**
Gender	Female (37), Male (32), Genderqueer/gender non-conforming (2), Transgender male (2), Transgender female (1)
Nationality	United States (61), Azerbaijan (2), China (2), Bangladesh (1), Colombia (1), India (1), Ireland (1), Kazakhstan (1), Philippines (1), Russia (1), Saudi Arabia (1), Vietnam (1)
Major	Math/engineering (35), Other (19), Natural Sciences (10), Social Sciences (10), Business (6), Humanities and arts (5), Communications (3)
School State	Massachusetts (24), Utah (15), New Jersey (12), Ohio (8), Pennsylvania (5), Texas (3), Tennessee (2), Wisconsin (2), Michigan (1)
School Type	Private (36), Public (36)

Study Modifications

One of the key differences between this study and our original study was that there was no Zoom component this time around and I did not directly interact with any of the participants. Instead, participants were allowed to fill out the Qualtrics survey on their own time. This experimental set-up more closely resembles how users might interact with the explainable outside of a lab setting.

Participants were also assigned randomly to one of our three explainable conditions (i.e., Long Limitations, Short Limitations, or No Limitations), making this a between-subjects experiment, and the post-test included an additional decision scenario component. In our new survey, we added some additional demographic questions in our pre-test to learn more about our subject pool (e.g., nationality, school location, public/private institution, etc.). Additionally, a new self-efficacy question was included to probe prior participants' experience with ASL; this question was based on [106].

In our pre-test, we decided to add another attention check, which was adapted from [40] as done in [63]. This adheres to the recommendation from [4] of including at least two attention checks in online surveys. Our other attention check can be found in the post-test section targeting participant mental models of BKT (see Question 5a in Appendix B.2.1); this was included in our original study as well. Like [63], we immediately disqualified participants who failed the first attention check in our pre-test and omitted responses from those who failed the second attention check in our post-test. All of these modifications to our survey are illustrated in Appendix B.1.

Participant Recruitment

For this study, we recruited undergraduates from various colleges and universities across the United States. Because all of the participants in our previous work were from Williams College, a small, rural liberal arts college, we were aiming for a more diverse and varied subject pool for this study. Thus, Prof. Howley and I sent out recruitment emails to several private and public institutions ranging in size, location, and other factors. We asked our contacts at each institution to distribute

our recruitment message to as many undergraduates as possible (e.g., posting on "Daily Messages" or reaching out to various undergraduate groups/organizations). Students were first asked to fill out a Google Form to indicate their interest in participating in our study. After collecting sign ups, we sent participants the link to our Qualtrics survey in a separate email. Participants were then given approximately one week to finish the survey and instructed to complete it in one sitting.

We ended up sending the survey to 197 undergraduates, 128 of whom responded. We disqualified 54 of these students for failing to pass our first attention check, resulting in a sample size of 74. We later omitted 10 additional participants, who completed the survey multiple times and/or were designated as outliers due to abnormally long survey times. All the remaining participants passed our second attention check. Some demographic information about our participants—prior to the second round of disqualification—is shown in Table 6.2. Some of these counts may not sum up perfectly to 74 because of participant non-response; for major, specifically, students were allowed to select more than one option. Again, none of our participants had prior experience with BKT, but one mentioned some familiarity with ASL.

Many recent studies looking at XAI methods and user perceptions of algorithms have opted to crowdsource participants from online marketplaces such as Amazon Mechanical Turk (MTurk) [14, 22, 63, 89]. However, we recruited college students because they represent our target users. After all, my explainable is the most relevant to people who might actually encounter BKT or similar learning analytics systems, which frequently come up in undergraduate courses [17, 53]. MTurk workers also have questionable external validity, considering their typical demographics [54] and rather low failure rates on attention checks [48]. Additionally, previous research suggests that crowd workers may already view algorithms more favorably than the general public as a whole [110], which could confound our results.

6.3 Results & Discussion

In this section, I present and discuss the preliminary results from our exploratory data analysis. Further analysis will be conducted at a later date.

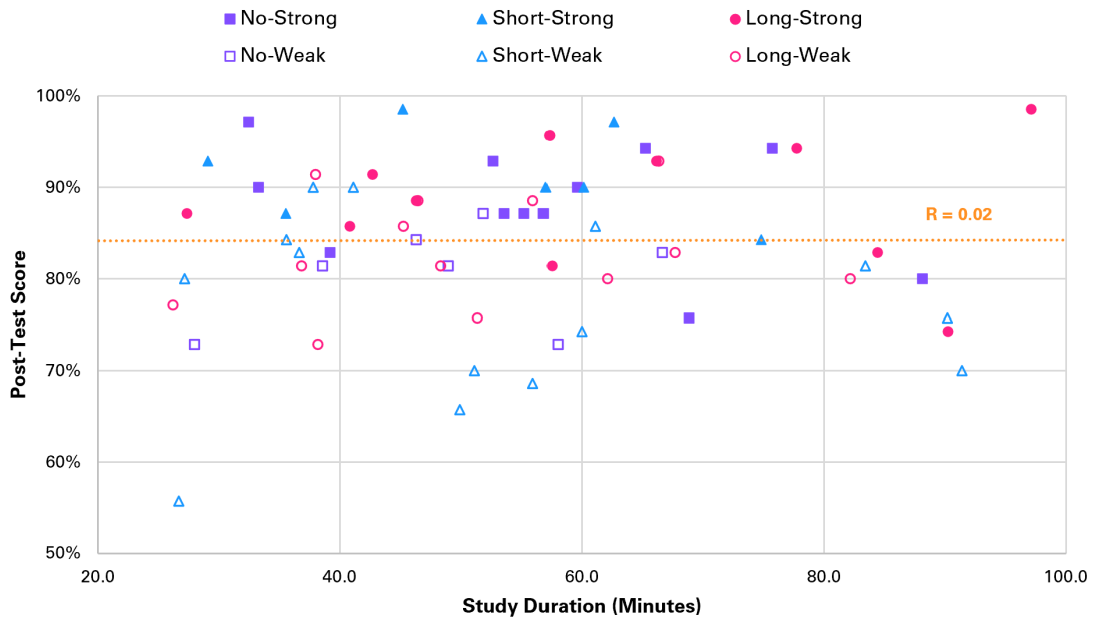
6.3.1 Are the Results From Our First Study Generalizable to Other College Students?

One of our primary goals with this additional work was to investigate whether the results obtained from our first study (Chapter 5) were replicable and generalizable to other college students outside of Williams.

Time-On-Task & Math/CS Background

We first decided to look at the impact of participant time-on-task and math/CS background on learning, as done previously in Section 5.2, through Figure 6.2. In this graph, participants in the *no limitations* condition are denoted in purple, participants in the *short limitations* condition are denoted in blue, and participants in the *long limitations* condition are denoted in pink. Additionally,

Figure 6.2: Participant Post-Test Scores vs. Study Duration

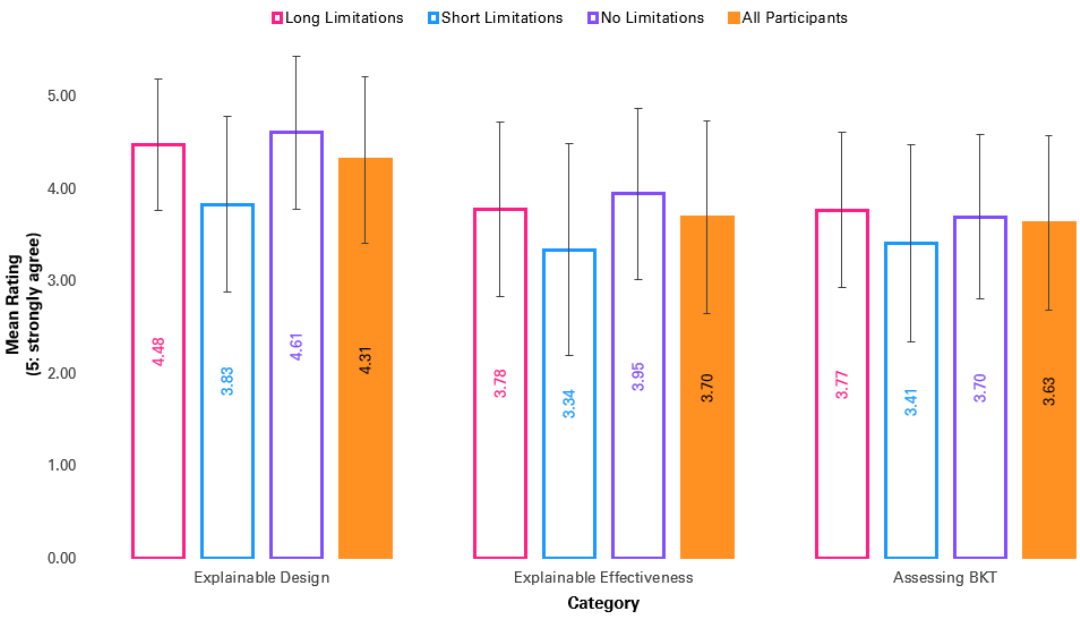


participants with stronger math/CS backgrounds are represented by solid markers (e.g., ●), while participants with weaker math/CS backgrounds are represented by outlined markers (e.g., ○). Out of our final 64 participants, 24 were randomly assigned to the *long limitations condition*, 21 were randomly assigned to the *short limitations condition*, and 19 were randomly assigned to the *no limitations condition*. 30 participants were identified as having stronger math/CS proficiency, and 34 participants were identified as having weaker math/CS proficiency.

As shown in Figure 6.2, we no longer have a clear correlation between study duration and post-test score (R = 0.02), and the reported times seem almost uniformly distributed. However, this is likely due to the fact that each explainable version is designed to take differing amounts of time, and variable time effects were expected as we wanted to see if differences in learning would arise from different limitation conditions. As expected, the *no limitations* condition was the shortest (average duration = 50.9 minutes), the *long limitations* condition was the longest (average duration = 56.4 minutes), and the *short limitations* condition was in between (average duration = 53.0 minutes). Looking at each condition separately may yield a more linear relationship. Another factor to consider is that although participants were instructed to complete the survey in one sitting, because I was not directly monitoring them, they may have taken breaks in between, potentially leading to longer reported times. It is interesting that the times between conditions did not seem to differ as drastically as we initially thought, which could suggest that participants tend to spend a similar time interacting with the explainable, regardless of how much information is presented.

Regarding user backgrounds, there also does not seem to be as strong of a correlation between math/CS proficiency and time-on-task as previously witnessed. In fact, we saw that students in the stronger math/CS group finished the survey slower on average compared to students in the weaker

Figure 6.3: Participant Post-Test Ratings of Explainable Design, Explainable Effectiveness, and BKT Trust



math/CS group (strong average: 56.3 minutes vs. weak average: 50.4 minutes). Perhaps students with stronger math/CS backgrounds engaged more deeply with the material and thus took longer to complete our study; it is also possible that these students happened to spend more time answering the post-test questions or explaining their answers.

One result we were able to replicate though was that students in the stronger math/CS group did achieve higher post-test scores than their peers on average (strong average: 88.9% accuracy vs. weak average: 81.9% accuracy). However, again, prior math/CS proficiency does not seem to be a prerequisite to successful learning with any of our explainable versions, as there were students in both the strong/weak groups who performed strongly on our post-test questions targeting the KCs of BKT (Figure 6.2).

Participant Ratings

Next, we looked at how participant ratings on our pre- and post-tests compared to our original study. As illustrated in Figure 6.3, which denotes mean ratings with standard deviation error bars, participants overall enjoyed the our **explainable design**, and this was the category that received the highest overall ratings across all three conditions. With **explainable effectiveness**, the ratings were a bit lower, mirroring the trend we observed previously; this is understandable, however, again due to the inclusion of the statements, "I could generally explain how BKT works to another person" and "I feel comfortable using BKT systems after completing this explainable." As was the case last time, it is likely that our students did not rate these statements as highly due to this study being their first exposure to BKT.

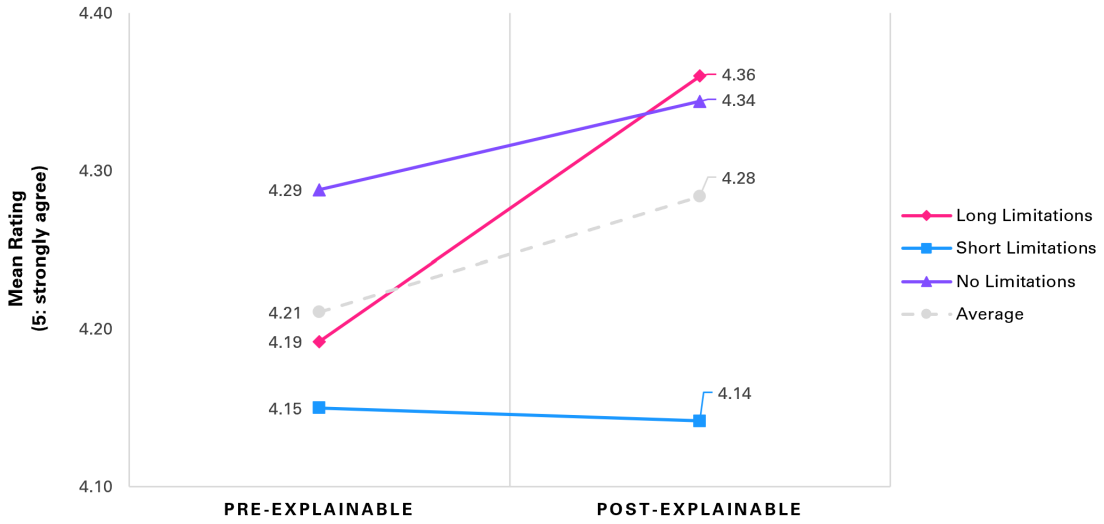
Interestingly, participants in the *no limitations* condition gave the highest ratings in both these categories, which perhaps offers evidence in favor of a simpler, less time-consuming design. On the other hand, participants in the *short limitations* condition gave the lowest ratings in both these categories. This might demonstrate a dissatisfaction with the amount of information presented about BKT's limitations in this explainable version, or a dislike of the lack of interactive components when teaching about these potential flaws and biases. One participant in this condition commented, "There was a lot of reading. I read through all of it but I forgot some of the things I was told. [If] I was given a video or a read a long- I think it would help me better retain information." Others from the *short limitations* group also mentioned how "sometimes it was wordy or more complex than I could follow." Regarding explainable effectiveness, several participants shared their desire to have more review before being tested on BKT concepts, which is something to keep in mind for future explainables: "I really liked the explainable, but once I closed it I had a hard time remembering each of the parameters and their meanings. I think if there was one more round of examples or maybe if the relationship between the name of the parameter and its meaning was pointed out one more time I would have remembered it better."

The **assessing BKT** category received the lowest overall ratings out of the three, with participants in the *no* and *long limitations* conditions having similar levels of trust in the algorithm, and participants in the *short limitations* condition having the lowest (Figure 6.3). It was surprising that trust decreased when provided with some information about BKT's limitations, but increased again when provided with additional information. This could be related to my hypothesis earlier about dissatisfaction with certain levels of algorithmic transparency, as many participants in the *short limitations* condition expressed an interest in learning more about BKT before assessing the algorithm's capabilities, stating, "[I] don't feel like I know enough and retained enough from the explainable to judge BKT" and "I think BKT is really interesting, but [I] would want to learn more about it and hear about how it is being used in different settings before I would have more trust."

Finally, we examined changes in participant **attitudes toward AI** more broadly from before and completing the explainable, as shown in Figure 6.4. Similar to our first study, a slight increase in ratings was observed between pre- and post-explainable ratings (pre: 4.21 vs. post: 4.28). But if we break down these ratings by condition, we see that while ratings in the *no* and *long limitations* conditions increased post-explainable, ratings in the *short limitations* conditions stayed relatively constant and even decreased slightly. Again, it is possible that participants who saw the explainable with *short limitations* did not receive enough information about BKT's potential flaws to significantly update their view of AI. This also suggests that it is potentially better to leave out information than not do it justice in post-hoc explainables.

One participant in the *no limitations* condition wrote, "My opinion is pretty similar to as it was before. However, I think I am now a little bit more welcoming of AI due to the application of BKT and how this could efficiently and potentially more accurate evaluate someone's performance," demonstrating how generally, our explainable increased optimism about the future of AI. Similarly, another participant stated, "I am very excited to learn more about AI systems and hope to be able to build some myself one day." Just like our previous results, we also saw evidence of continued doubts about the increasing prevalence of AI in society. Several students mentioned the problem

Figure 6.4: Changes in Average Participant Attitudes Toward AI



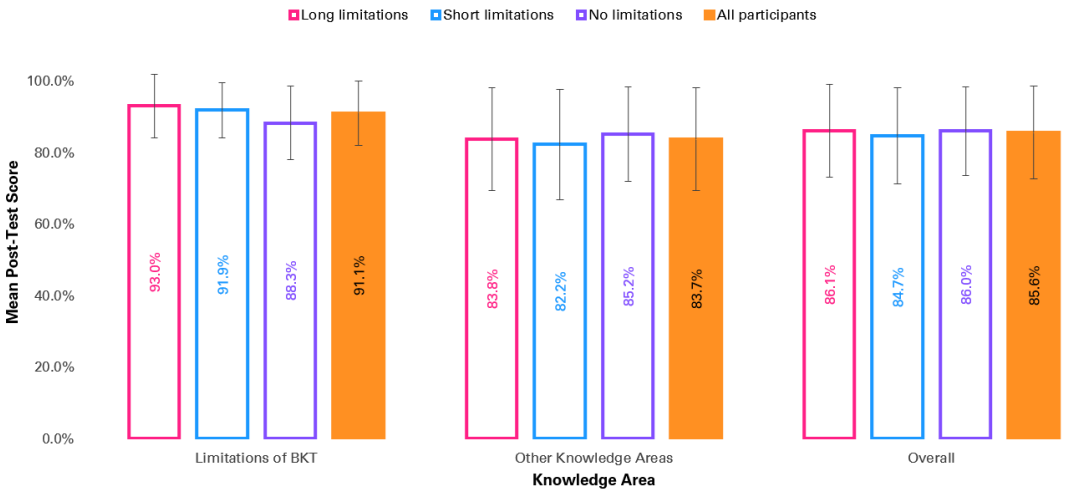
of biased data, indicating a nuanced understanding of the relationship between humans and AI algorithms, writing how "humans code AI... If the human is biased, they code that bias, creating a giant discriminatory machine" and "I don't believe there's anything inherently harmful about AI systems/algorithms because that would be a harsh generalization. It's mostly the creator's intentions and choices they make in creating its functions that I am concerned about." Other participants were also interested in learning more about the ethics of AI, suggesting another possible addition to future XAI systems.

6.3.2 How Does Decreasing the Transparency of BKT's Limitations Affect Algorithmic Understanding?

Next, we investigated the precise effect of varying algorithmic transparency on user understanding of BKT through participant post-test scores. Figure 6.5 breaks down participant post-test scores by knowledge area and limitation condition. Again, each bar denotes an average accuracy, and the error bars represent the corresponding standard deviation. We decided to isolate **Limitations of BKT** from our other three knowledge areas since this study specifically aimed to determine how the amount of information provided about BKT's limitations would impact user outcomes.

Overall, participants in all three conditions performed similarly, achieving post-test accuracies of around 85-86%, which is comparable to the results from our previous study (Figure 5.1). We were also surprised to see that outside of the **Limitations of BKT** knowledge area, participants in the *no limitations* condition actually achieved the highest average post-test scores overall. This observation could be potentially explained by time-on-task [97], considering that these students spent similar amounts of time on the study as students in the other two conditions, because that would mean they could spend more time on average learning fewer concepts, since there was less content in the

Figure 6.5: Participant Post-Test Scores by Knowledge Area



explainable with *no limitations*. Likewise, students in the *no limitations* condition likely experienced less cognitive overload than their peers [87], so they were possibly able to remember the information they did learn better on the post-test. It is also unsurprising that the *no limitations* group performed the worst on questions pertaining to **Limitations of BKT** as they did not receive any information about the algorithm's potential flaws and biases. One common misunderstanding was that many of these students thought that time could be accounted for by BKT, likely due to confusion with speed of signing being associated with **P(transit)** in our explainable (Figure 4.7).

Regardless, students in the *no limitations* condition still demonstrated a strong understanding of BKT's limitations, considering that they achieved an average score of 88.3% on this section of the post-test (Figure 6.5). This result evidences again how non-experts may have more intuition about algorithmic limitations than expected. As a whole, participants scored an impressive 91.1% accuracy on average for **Limitations of BKT**. We were again impressed by our participants' abilities to apply knowledge from our explainable throughout this section of the post-test to infer information about BKT's potential shortcomings; for example, some participants also mentioned that location could impact student learning, but is not accounted for by BKT: "Environments can affect how well someone can learn/master a skill... [but] there is no place for location to be considered a factor in BKT." Notably, this was the only BKT knowledge area where participants in the *long limitations* section performed the best, but this does offer evidence in favor of the importance of including ample information about algorithmic flaws and biases in post-hoc explainables.

In this study, **Identifying Priors** was again the knowledge area that participants struggled the most with (average score: 66.8%), especially regarding remembering the names and definitions of the four BKT parameters. Most students could not recall the acceptable bounds for **P(slip)** and **P(guess)** either, which again strengthens our findings from before and evidences the potential need for more emphasis on teaching foundational concepts in AI explainables. Regarding the **Identifying Changed Parameters** knowledge area (mean accuracy: 88.7%), the main problem we identified was

that some participants thought **P(init)** never changes (e.g., "The init value stays the same because it signifies existing knowledge prior to the assessment" and even simply "P(init) is constant"), so perhaps our explanation did not make the dynamic nature of **P(init)** clear enough. This could also be reflective of our participants' beliefs in growth vs. fixed mindsets [38], as those who embody more fixed mindsets may think people are either good at a skill or not (e.g., math), and thus **P(init)** would always remain at the same value.

Conversely, **Evaluating P(init)** seemed to be the most intuitive knowledge area for our participants this time around, with an average post-test score of 95.8% across all three conditions. On these questions, students demonstrated a strong understanding of different factors that could impact the reliability of BKT's estimates of **P(init)**. The most common concept they struggled with was understanding that the threshold for mastery can vary. Most participants were loathe to deviate from the typical value of 0.95 [27], which we also observed in our original study, commenting how, "I don't see why the typical threshold for mastery should change based on the situation." This hesitance to stray from 0.95 may indicate a need to emphasize this topic more in our explainable, but it could potentially also be the result of students' theory of knowledge and the difficulty of communicating nuance and uncertainty [65].

For those who were willing to assign different thresholds depending on the scenario, it was interesting to see that many students believed the mastery threshold for medical school should be lower than a math worksheet (Question 6 in Appendix B.2.1). One student wrote, "Medical school is designed to teach people who aren't familiar with medicine so expecting a high level of pre-existing mastery defeats the purpose. With that in mind, any low-ish score should be acceptable." Other participants echoed these sentiments in various ways: "Since this is an acceptance exam, there should not be as significant of a presumption of knowledge." Previously, we presumed people would assign higher mastery thresholds to more "high-stakes" scenarios, but our participants offer compelling explanations as to why this might not always be the case.

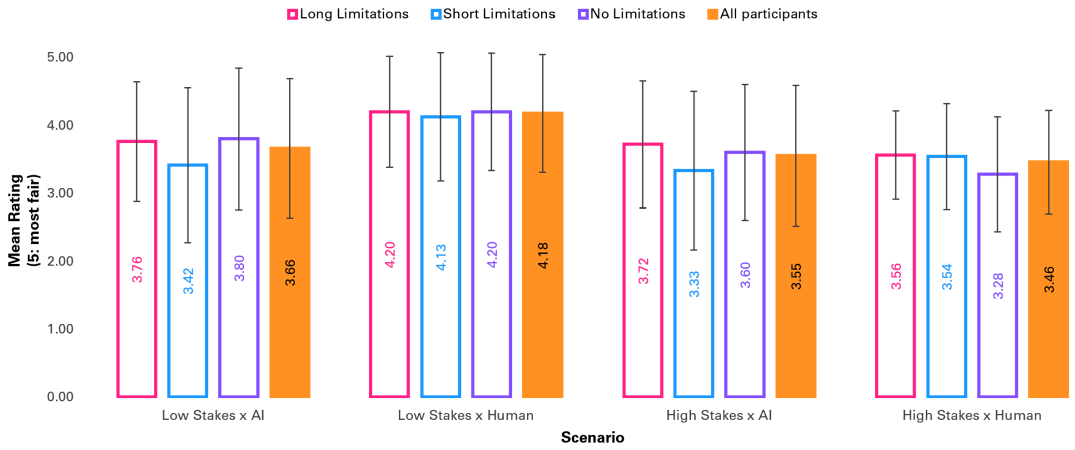
6.3.3 How Does Decreasing the Transparency of BKT's Limitations Affect Perceptions of Algorithmic Fairness and Trust?

Finally, we looked at how our different explainable conditions affected user perceptions of algorithmic fairness and trust via the decision scenarios we constructed (Table 6.1).

Perceptions of Fairness

Regarding fairness, we thought it was interesting that participants in the *long limitations* condition tended to assign the highest fairness ratings and participants in the *short limitations* condition tended to assign the lowest fairness ratings to our decision scenarios, illustrated in Figure 6.6 (error bars denote standard deviation). Again, the latter group might have been a little more skeptical of the AI decision-makers if they were not satisfied with the amount or type of information that was provided about BKT's limitations. It is also possible that providing more comprehensive information about BKT's limitations leads people to realize that algorithms are more fair than they had previously imagined, thus resulting in higher fairness ratings than the *short limitations* condition.

Figure 6.6: Participant Fairness Ratings by Scenario Type



And participants in the *no limitations* condition are likely working with their original perceptions of algorithmic fairness, which appear to be relatively high as well, since they did not receive any related information while completing our BKT explainable.

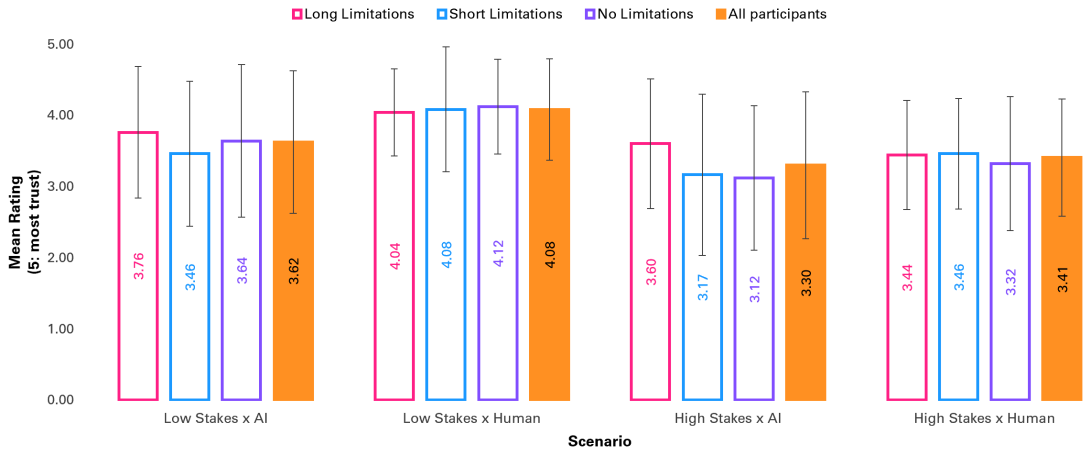
Additionally, we discovered that participants thought the decision was most fair in the low-stakes scenario involving a human decision-maker (average fairness rating: 4.18) and the least fair in the high-stakes scenario involving a human decision-maker (average fairness rating: 3.46). We were surprised to see that in high-stakes scenarios, the AI decision-maker was seen as more fair than the human decision-maker, but many participants used the subjectivity of human graders to explain their ratings. For example, some students described how with an AI algorithm, "the same criteria will be applied to everyone, so" it "would be very fair because it eliminates the internal bias that comes with a human person grading this test." Along with bias, participants cited the fact that "the AI algorithm is less likely to make an error compared to a human" as a reason why their fairness rating for the human grader was lower.

Many participants still had concerns about the fairness of these AI decisions, however. Students demonstrated sophisticated understanding that different algorithms may come with certain strengths and weaknesses: "Depends on the style of the text. I think that an algorithm could perfectly handle a multiple-choice test, but may struggle to accurate judge the knowledge of a student's written work." Some also mentioned that "although accuracy is very important, what [students] need now is encouragement," emphasizing the lack of "humanness" of algorithmic decision-makers.

Perceptions of Trust

Figure 6.7 (error bars denote standard deviation) demonstrates that as a whole, our participants assigned slightly lower trust ratings to each decision scenario compared with fairness ratings. Intuitively, this makes sense because even if you perceive a decision to be fair, you still might not completely trust it. We also see that in this graph, there is no participant condition that consis-

Figure 6.7: Participant Trust Ratings by Scenario Type



tently is more or less trusting of these decisions than the others. However, it does appear that ratings across conditions are more consistent when the scenario involves a human decision-maker, rather than an AI decision-maker, with the least variation occurring in the low-stakes, human decision-maker scenario. A similar trend can also be witnessed in Figure 6.6.

Overall, it seems that participants in the *long limitations* condition assigned higher ratings to the scenarios involving an AI decision-maker, suggesting that in this case, algorithmic transparency does seem to increase trust, although it is debatable whether this is a positive or negative user outcome. As previous research has shown, increased trust can easily lead to overreliance on AI models and systems [22, 89].

Mirroring our fairness ratings, participants tended to trust the decision most in the low-stakes scenario involving a human decision-maker (average trust rating: 4.08). But this time, they trusted the decision least in the high-stakes scenario involving an AI decision-maker (average trust rating: 3.30). So although AI algorithms might be perceived to be more fair than humans in high-stakes scenarios, people might still trust human decisions more. Participants were quick to point out how potential shortcomings of AI decision-makers could decrease trust, noting that "the AI may or may not be able to tell if Mars truly understands the material or if they copied someone else's answers" and "first graders do not have the best handwriting so it may be difficult for an algorithm to understand the writing." Others stated how in general, "Without any human oversight, I worry about trusting an AI." On the other hand, some participants likened the limitations of algorithms to humans: "This algorithm seems to make sense to me and while it does have some flaws so do people." Students also expressed that an algorithm's trustworthiness might depend on many factors, but "as long as the test is based on logic" or "the AI is... trained for the answers on this specific test [it] is trustworthy."

Regarding trust in human decision-makers, most participants agreed that a human "is more likely to understand if [a student] did something that was nearly correct than an AI might." Many students also mentioned the fact that in the case of a teacher, for example, they have "been trained to

properly evaluate students," whereas an algorithm might not have. But in the high stakes scenario, the concern of bias continued to prevade participant answers, indicating some distrust in humans as well. For instance, participants expressed how "a human is not fully objective" and "we do not know if this human grader is qualified. We also do not know if the human grader possesses some bias and shows preference towards some kinds of answers." Participants also had conflicting viewpoints on people's intentions; while some stated, "Humans are very bad at make decisions about determining other success. We know this from studies," others reported, "I have trust in humans... I think most people want whats best for others."

6.4 Limitations

This study shares many of the same limitations as our previous work (Section 5.4). For instance, although we expanded the subject pool beyond Williams College students, we did still target solely undergraduate students in the United States, so the external validity of our research may be limited. Additionally, we attempted to recruit participants from a variety of schools, but there could be some selection bias because Prof. Howley and I only reached out to institutions where we had contacts, and our sample is certainly not representative of all school types or locations. The majority of our students also self-reported their major as math/engineering, which may have skewed our results. However, BKT systems are most frequently used in math and statistics contexts [17, 53], so it seems reasonable to target students who are more likely to encounter the algorithm.

Our sample size was larger in this study (n = 64), but we would likely need more than ∼20 participants per condition to statistically confirm confidence in our findings. Additionally, it is possible that although we randomly assigned participants to explainable groups, confounding variables could still be affecting the results. For example, maybe participants who completed *long limitations* explainable are just more trusting people in general, so their trust ratings of our decision scenarios might not be higher solely due to explainable condition. We also only tested one decision scenario of each type, so our results regarding user perceptions of algorithmic fairness and trust may not necessarily be generalizable. At this point, we have not conducted official statistical tests to determine causality or significance, so for now, these results should again just be taken as correlational trends.

The same comments related to the criticisms of post-hoc explainables discussed previously apply here as well. Another challenge along these lines is determining the correct level of algorithmic transparency to include in each AI explanation. As every algorithm is inherently different in nature and may be used in different contexts, we will also need to determine a more systematic way of deciding how much information to provide end users about AI systems. It is understandably not sustainable—in terms of both time and resources–to conduct an experiment for each algorithm as I have done in this thesis.

6.5 Summary

In this study, we examined how algorithmic transparency affects user outcomes such as perceived fairness and trust. To do this, we created three different versions of our explainable, varying the

amount of information present about BKT's limitations, and added decision scenarios inspired by [63] to our post-test to measure user perceptions of algorithmic fairness and trust. The results from our user study demonstrate that *short limitations* condition led to lower overall ratings regarding explainable effectiveness/design and BKT trust, while the *no limitations* condition led to the highest overall ratings in these categories. Our explainable also increased confidence in the future of AI for all participants, except for those in the *short limitations* group.

Participants in all three conditions performed similarly overall on our post-test, receiving the highest scores on questions pertaining to **Evaluating P(init)** and the lowest scores on questions pertaining to **Identifying Priors**. Students in the *no limitations* condition achieved the highest average accuracy over all knowledge areas except **Limitations of BKT**, where students in the *long limitations* condition performed the best. This evidences that including information about algorithmic limitations in XAI systems is still beneficial and advisable. Finally, we discovered that in low-stakes scenarios, participants assign higher fairness and trust ratings to human decision-makers over AI. But in high-stakes scenarios, human decision-makers are perceived as less fair but more trustworthy. In general, participants in the *short limitations* group tended to assign lower fairness ratings. On the other hand, participants in the *long limitations* group assigned the highest fairness ratings overall and were more likely to trust decisions made by algorithms. Thus, it seems that more algorithmic transparency in our BKT explainable increased user perceptions of fairness and trust overall.

However, as before, these results may have certain limitations as discussed in the preceding section. Next steps for this study include completing a more comprehensive analysis of our data.

Chapter 7

Conclusion

7.1 Contributions

In my thesis, I successfully address the following research questions introduced in Chapters 1 by conducting two studies:

- **Study 1:**
  - What are the knowledge components of algorithmic understanding for BKT?
  - Can we build an explainable to teach these BKT concepts?
  - What factors impact successful learning with our explainable?
  - How does our explainable impact user attitudes toward BKT and AI more broadly?
- **Study 2:**
  - Are the results from our first study generalizable to other college students?
  - How does decreasing the transparency of BKT's limitations affect algorithmic understanding?
  - How does decreasing the transparency of BKT's limitations affect perceptions of algorithmic fairness and trust?

In our first study, I found Cognitive Task Analysis to be a productive method toward identifying concrete steps that experts take to conceptualize Bayesian Knowledge Tracing, an AI algorithm. By analyzing expert responses to a variety of BKT vignettes, we identified four main knowledge areas to consider when explaining this algorithm: (1) **Identifying Priors**, (2) **Identifying Changed Parameters**, (3) **Evaluating P(init)**, and (4) **Limitations of BKT**. These four knowledge areas were then further split into 19 individual knowledge components described in Section 3.2 that represent the generalized steps experts took to respond to the posed scenarios.

Afterwards, I developed a post-hoc explainable to impart BKT concepts to representative end users using best practices from user-centered design and teaching & learning theory. Following the

principles of Backward Design, we targeted specific subsets of the identified KCs in our explainable, evaluated the effectiveness of instructional content along those subsets via user studies, and measured the impact of algorithmic understanding on behaviors of non-expert student users of BKT systems. Our results demonstrate the effectiveness of this learning sciences inspired approach for XAI, as all participants achieved high post-test scores and demonstrated sound understanding of BKT's potential flaws and biases. The **Identifying Priors** knowledge area was the most challenging for students to grasp, while performance was the highest for questions pertaining to **Identifying Changed Priors**. We also discovered that student comfort levels with math/CS may impact the amount of time they take to complete the explainable, but prior expertise is not a prerequisite to successful learning. My explainable increased confidence in the future of AI as well, evidencing the impact of algorithmic understanding on user attitudes.

In our second study, we implemented additional versions of my explainable and created decision scenarios to explore more deeply the effect of algorithmic transparency on user perceptions of fairness and trust. We conducted another round of user studies, successfully replicating the larger trends regarding participant attitudes towards our explainable, BKT, and AI from our first study. Participants across all three conditions performed similarly well on our post-test, although certain conditions did achieve higher scores in certain knowledge areas. In this study, **Identifying Priors** was again the most challenging knowledge area for students to grasp, but this time, **Evaluating P(init)** yielded the highest post-test scores. We also observed that participants in the *short limitations* condition perceived decisions to be less fair on average than participants in the *no* or *long limitations* conditions for both AI and human decision-makers, while participants in the *long limitations* condition were the most likely to perceive decisions made by algorithms as trustworthy.

These findings can 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 as proposed in Figure 1.1. 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 novice acquisition and shifts in decision-making patterns connected to each KC provides a generalizable framework (Figure 1.2) for building evidence-based post-hoc AI explanations that are accessible even to non-AI/ML experts. For educational applications in particular, our approach can also help enhance the accessibility and efficacy of algorithmically augmented learning environments for students and teachers, which will be especially pertinent moving forward as novel forms of remote and hybrid learning emerge in the imminent post-pandemic future.

7.2 Future Work

7.2.1 Generalizing Our Approach

Now that that we have successfully applied CTA and pedagogical theory to design a student-centered, post-hoc explainable for BKT, next steps include investigating the wider applicability of this method. For instance, in this study, we target student users of BKT, but students are not the only users of these learning analytics systems. Teachers are important stakeholders as well [110], so future work

should consider repeating our CTA protocol with instructors. Then, we can compare the KCs derived from teachers with our original set of KCs from student experts and determine if and what differences exist. This information can be used to decide whether different explainables should be constructed for different stakeholders, or if a more general AI explanation could suffice, given that user goals are sufficiently aligned. The results from this work could also be used to inform other XAI applications where multiple stakeholders are involved. Participant feedback, as discussed in Section 5.1.2, can also be applied to improve the design and efficacy of future explainables.

Once we finish analyzing the data, it may also be of interest to repeat our second study (Chapter 6) with students who have more diverse backgrounds (e.g., all humanities majors at a single institution). This could help us determine whether our results are truly generalizable to those who may have less experience and confidence with math and CS. Additionally, it would be worthwhile to try applying our novel, evidence-based framework (summarized in Figure 1.2) to algorithms and domains outside of education to verify the robustness and generalizability of our approach. For instance, can these methods from the learning sciences and human-computer interaction be used to design effective, user-centered explanations for high-stakes applications of AI in healthcare and criminal justice systems? As mentioned in Section 5.4, future work should also investigate how to most productively deploy post-hoc AI explanations in industry and other real-world settings.

7.2.2 Assessing Behavioral Outcomes

Another direction to pursue is further studying the impact of XAI systems on user outcomes such as trust and fairness (Figure 1.1), building off the work from my second study discussed in Chapter 6. After all, increased understanding is not the only goal of XAI. Trust and informed decision-making are additional desirable outcomes in the Fairness, Accountability, and Transparency (FAccT) machine learning model [1].

But to truly confirm external validity, behavioral measures must be developed to define how changes in user behavior can be observed. For example, asking participants to gamble or bet on particular outcomes given limited resources [46] is a common approach within the decision sciences. In educational settings, participants may be left alone for brief unstructured time after an intervention to allow voluntary engagement with the intervention, which is then taken as a behavioral measure of interest [16]. As a first step in this direction, self-reported decisions can feasibly be coupled with behavioral measures to provide more concrete evidence of how user perceptions influence visible behavior with AI systems such as BKT.

Ultimately, the goal would be to see whether user understanding changes the way people interact with algorithms or the decisions they make while using AI/ML systems. For example, it maybe be possible to measure Human-AI task performance in learning settings after viewing different post-hoc explanations, taking inspiration from [14, 22]. To continue my work with BKT, it would be exciting to have the opportunity to partner with a real learning analytics system that uses this algorithm, such as the *Open Analytics Research Service* [17] or *Lynette* [53]. Then, we could examine how increasing algorithmic understanding in the field influences real decisions made by instructors and students impacted by BKT systems.

7.3 Summary

The goal of my thesis was to design a more robust, evidence-based framework for creating and evaluating post-hoc XAI. To do this, we applied methods from the learning sciences and HCI such as Cognitive Task Analysis and Backward Design to identify the necessary knowledge components to teach stakeholders about Bayesian Knowledge Tracing, our pilot algorithm, and design an explainable targeting these KCs. We developed assessments and ran a user study to evaluate the effectiveness of our BKT explainable, finding that it was successful in imparting a strong basic understanding of the algorithm to student non-experts. Our user study also revealed time-on-task and math/CS background as two factors that potentially effective user learning with explainables. Participants seemed more optimistic about the future of AI after engaging with our BKT explanation as well.

We then ran a second study to investigate more deeply how algorithmic transparency impacts user outcomes. By designing additional explainable versions and adding decision scenarios to our post-test, we were able to assess how varying the amount of information presented to users about BKT's limitations affects perceptions of algorithmic fairness and trust. This study revealed that including insufficient information in our explainable may cause user dissatisfaction and lower fairness ratings. However, more algorithmic transparency can lead to higher trust in AI decision-makers. These results suggest that our novel learning sciences inspired approach is effective in guiding the design of post-hoc AI explainables, and can potentially be generalized to other algorithms and contexts. Future work includes assessing the wider applicability of our framework and directly measuring changes in user behaviors that arise from algorithmic understanding.

Appendices

Appendix A

Cognitive Task Analysis Questions

This appendix contains the full list of vignette survey questions originally written by Noah Cowit '20 and posed to student BKT experts during our CTA protocol, as described in Section 3.1. Questions 11 and 12 were excluded from our final analysis as they targeted related topics to BKT such as skill mapping [20], but did not directly pertain to the algorithm itself. Understandably, our experts had more variable levels of knowledge about these subject areas, so the *empirical/prescriptive* CTA conducted in this study would likely not be able to accurately capture the relevant KCs.

Question 1

Sandy is a student who sleeps a lot. This often gets in the way of his studies. Sandy has a test in his geosciences class coming up. He has often missed classes in geoscience because of his mid-day naps. The test consists of 5-option multiple choice questions, with no penalties for incorrect answers. Sandy is given immediate feedback regarding the correctness of his answers. The test answers derive from a lab activity, and answers build on each other. Despite his sleep schedule, Sandy is a reasonably studious person.

(a) What do you think are reasonable values for the BKT parameters at the beginning of this test? Please talk me through your reasoning.

- **P(init):**
- **P(transit):**
- **P(guess):**
- **P(slip):**

(b) Sandy gets the first answer correct. What are reasonable values for the BKT parameters now? Please talk me through your reasoning.

- **P(init):**
- **P(transit):**

- **P(guess):**
- **P(slip):**

(c) Sandy gets 4 of the final 5 questions of the test correct. What do you think is a reasonable value of **P(init)** at the end of the test? Explain your response.

(d) Based on your answer to the previous question, would you say Sandy has "mastered" the material? Why or why not?

Question 2

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 format, where the user must match 10 words to their definitions.

(a) What do you think are reasonable values for the BKT parameters at the beginning of one of these vocab tests? Please talk me through your reasoning.

- **P(init):**
- **P(transit):**
- **P(guess):**
- **P(slip):**

(b) Amari got 6 out of 10 questions correct on their test. At the end of the test, BKT suggests Amari has not mastered the material. What is your interpretation of this analysis?

Question 3

Margo is a typical student who enjoys studying history. Like many students, Margo can learn by answering questions, seeing what they get wrong, and learning from their mistakes. Margo is taking a test in which questions are related, but there is *no immediate feedback* given as to the correctness of their answers.

(a) What is a reasonable value for **P(transit)** in this scenario?

(b) Consider if Margo was given the correct answers in real time, but no explanation. What would be a reasonable value of **P(transit)** now?

(c) What is a reasonable value of **P(transit)** if Margo was given the answers and an explanation?

Question 4

Darby likes to read ahead in her math class until she is sure she has learned the material. One day, Darby is asked to do a worksheet on material she has already gone through.

(a) Is 0.4 a reasonable value for **P(init)** in this situation? Explain.

(b) Darby gets the first question incorrect, a simple fact mentioned multiple times in her textbook. How would *you* classify this incorrect answer in the context of BKT?

Question 5

Marc has just returned to school after a summer vacation. Her teacher gives her a handout with review problems. Marc cannot remember how to add fractions with different denominators for her life. Marc is upset. She got a very good grade on the test for the material last year.

(a) If this previous test were evaluated by BKT, what would be a reasonable value for **P(init)** *before* Marc attempts the review problems?

(b) Describe this behavior of BKT and why it is or is not intuitive/sensible.

Question 6

Rory loves math but is afraid of calculators. They use mental math whenever they can, even on tests, despite the fact that they are not consistently successful. When teachers correct their tests, they find that Rory knows exactly what they are doing, but they are just making minor mental math mistakes (i.e. forgetting to carry to the next column, etc.).

(a) Rory is taking their final exam, which is scored through BKT. Their final grade depends completely on the correctness of their final answers. The exam is 3 questions long, with each question having significant calculative complexity. Talk me through what you think a reasonable value of **P(init)** is after the completion of the exam.

(b) What if Rory was taking the same test, except each question was divided into 8 "sub-questions," which were graded based on logical correctness from the previous sub-question. Would Rory's **P(init)** be likely to change? How and why?

Question 7

Pinto has just taken a very odd exam. It was a 3-option multiple choice test, and every question had two correct answers. Additionally, Pinto drank too much coffee, and found it difficult to fill in the bubbles correctly on the answer sheet (i.e. He fills in "C" when he wants to select "A").

(a) What do you think are reasonable values for **P(slip)** and **P(guess)** at the beginning of this test? Please talk me through your reasoning.

- **P(guess):**
- **P(slip):**

(b) After Pinto completes the test, his **P(init)** is 0.55. Please interpret this result.

Question 8

Kim recently took an extensive test on Nuclear Power Plant operations. The test was a standard 4-option multiple choice exam, and Kim finished in a reasonable time frame. After taking the test, Kim's **P(init)** is 0.97. Kim lives near a nuclear power plant, and the operator is out sick for the day. The power plant desperately needs a temporary operator, and would rather not pay to helicopter one in. If the operator is not a master in Nuclear Power Plant operations, a nuclear meltdown is likely to occur. Should Kim be offered this job? Explain your response.

Question 9

Burt and Ernie both take the same test. Incredibly, they not only get the same score, but their answers are identical. Burt and Ernie took the test at the same time, in separate locations. No cheating occurred. However, while Burt took 17 minutes to take the test, Ernie took 4 hours. After the test, BKT gave Burt a **P(init)** of 0.99. What would be a reasonable value for Ernie's **P(init)** after the test, as given by BKT?

Question 10

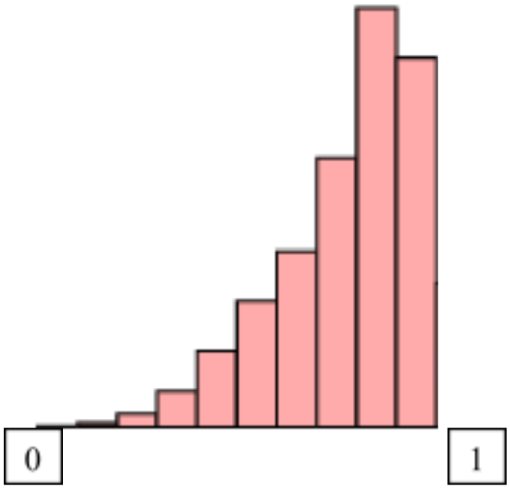
At University College, Jamie takes a physics test scored through BKT. They receive their results and see that their **P(init)** is 0.80.

(a) How should Jamie interpret this score?

(b) Jamie learns that the class average of **P(init)** for this physics test is 0.55, with a 75th percentile of 0.70. With the curve, they are given an A. Should this change Jamie's interpretation of their score? Why and in what way?

Question 11

Mr. Sandman is a substitute teacher at a local public high school. The regular teacher is lazy and asks him to grade a recent exam his students took. The test was taken online. When Mr. Sandman looks at the scores, he sees that they are assessed by BKT.



(a) Assuming that the scores range between 0 and 1 on the **P(init)** spectrum, and a near majority of students have achieved mastery (see figure above), what would be a reasonable grade for a student whose **P(init)** is 0.7, on an A-F scale?

(b) What about a student with a **P(init)** value of 0.91?

Question 12

Professor Plum receives the results of her class' ECON 110 exam. She sees the students generally did well on the analyzing GDP data section of the test. But there was one question that most students answered incorrectly. The section of the test is included on the next page.

(a) What could be a reason so many students got question 3 wrong? Explain your answer.

(b) What could be a way for Professor Plum to improve student performance on this question?

Data Analysis: This is a comparison of three countries, nation X, Y, and Z. Nation X has a GDP of 23 billion, and a population of 1 million. Nation Y has a GDP of 7 Trillion, and a population of 200 million. Nation Z has a GDP of 300 billion, and a population of 100 million.

1. Which nation has the lowest GDP?

   ○ Nation X
   ○ Nation Y
   ○ Nation Z

   Answer: Nation X. *Correct Answer Percentage:* 97%.

2. The company SeaTaxi wants to open up a new branch. They do not yet have a branch in Nation X, Y, or Z. They want to reach the largest market possible. Which nation should they open up a branch in?

   ○ Nation X
   ○ Nation Y
   ○ Nation Z

   Answer: Nation Y. *Correct Answer Percentage:* 93%.

3. Another company, Hoeing, has a different business model to SeaTaxi. Hoeing has seen their largest profit centers in markets where average income is greatest. Which Nation should Hoeing open a branch in?

   ○ Nation X
   ○ Nation Y
   ○ Nation Z

   Answer: Nation X. *Correct Answer Percentage:* 37%.

4. ...

Appendix B

Pre- & Post-Test Questions

Below, I will include the full list of pre- and post-test questions used to evaluate my final explainable, as described in Sections 4.3.1 and 6.2.

- Required questions are denoted with a *.
- Answer options for *single-answer multiple choice* questions are denoted with a ○.
- Answer options for *select all that apply* questions are denoted with a □.
- Questions without answer options are *open-response*.
- Questions added for our second study are denoted with blue.

B.1 Pre-Test

Pre-test questions mainly pertained to participant demographics (e.g., age, gender, education level, etc.) and backgrounds with math and CS.

B.1.1 Demographic Information

Instructions: Your answers are completely anonymous and will only be used to help us learn more about our participant population.

1. How old are you?*

   ○ < 18
   ○ 18-24
   ○ 25-34
   ○ 35-44
   ○ 45-54
   ○ 55-64
   ○ 65+

2. To which gender identity do you most identify?*

   ○ Male
   ○ Female
   ○ Transgender Male
   ○ Transgender Female
   ○ Genderqueer/gender non-conforming
   ○ Different identity (please state):

3. What is your nationality?*

4. What is the highest level of education you have completed? **Please select "some college" if you are an undergraduate student.***

   ○ Less than high school
   ○ High school graduate
   ○ Some college
   ○ 2 year degree
   ○ 4 year degree
   ○ Professional degree
   ○ Doctorate degree

5. What was/is your main field(s) of study?

   ☐ Natural sciences
   ☐ Math/engineering
   ☐ Education
   ☐ Humanities and the arts
   ☐ Social sciences
   ☐ Business
   ☐ Communications
   ☐ Policy
   ☐ Other:

6. Which country is your current college/university located in?

7. If your current college/university is in the United States, what state is it located in?

8. Do you attend a **public** or **private** college/university?

   ○ Public
   ○ Private

9. Which of the following best describes your current occupation?*

   ○ Student
   ○ Employed full-time
   ○ Employed part-time
   ○ Self-employed
   ○ Home maker
   ○ Retired
   ○ Unemployed
   ○ Other:

10. **(Attention Check)** This study requires you to voice your opinion using the scales below. It is important that you take the time to read all instructions and that you read questions carefully before you answer them. Previous research on preferences has found that some people do not

take the time to read everything that is displayed in the questionnaire. The questions below serve to test whether you actually take the time to do so. Therefore, if you read this, please answer 'somewhat agree' to the first statement, and select the option directly to the left of that for the second statement. Thank you for participating and taking the time to read all instructions.*

	Strongly disagree				Strongly agree
I would prefer to live in a large city rather than a small city.	○	○	○	○	○
I would prefer to live in a city with many cultural opportunities, even if the cost of living was higher.	○	○	○	○	○

11. **Please rate how strongly you agree/disagree with the following statement:** I can confidently understand and communicate with American Sign Language (ASL). *Please select "strongly disagre" if you have no experience with ASL.**

○ Strongly disagree ○ Somewhat disagree ○ Neither agree nor disagree ○ Somewhat agree ○ Strongly agree

B.1.2 Math/CS Background

Instructions: Please answer all these questions on your own, without looking them up or referring to external resources. Remember, we are not evaluating you.

1. If we flip a fair coin twice, what is the probability that we get heads both times?*

2. If we roll two dice, what is the probability that they sum up to 3?*

3. For the following statements, use the scale on the right to select the answer that best describes you (1: strongly disagree to 5: strongly agree).*

Math/CS Knowledge	Strongly disagree				Strongly agree
I am confident in my ability to learn and apply knowledge about math/statistics.	○	○	○	○	○
I am confident in my ability to learn and apply knowledge about computer science.	○	○	○	○	○
I am confident in my ability to learn and apply knowledge about artificial intelligence (AI).	○	○	○	○	○

General Attitudes toward AI	Strongly disagree				Strongly agree
I believe AI does more good than harm.	○	○	○	○	○
I support the increasing prevalence of AI in society.	○	○	○	○	○
I am excited about the future of AI.	○	○	○	○	○
It is important to educate the public on how AI systems work.	○	○	○	○	○
It is important to educate the public on the potential harms and limitations of AI systems.	○	○	○	○	○

4. Are you familiar with the term "Bayesian?"*

   ○ Yes ○ No

5. To the best of your ability (without looking it up!), please define Bayesian.

6. Have you ever heard of Bayesian Knowledge Tracing (BKT)?*

   ○ Yes ○ No

7. Please describe what you know about BKT below (without looking it up!).

8. Which of these online learning sites are you familiar with/have you used before?*

   ☐ None
   ☐ Khan Academy
   ☐ Coursera
   ☐ LinkedIn Learning
   ☐ Carnegie Mellon University's Open Learning Initiative (OLI)
   ☐ Stanford Online
   ☐ edX
   ☐ Udacity
   ☐ Udemy
   ☐ Youtube
   ☐ Other:

B.2 Post-Test

Post-test questions primarily targeted participant mental models of BKT, explainable usability/user satisfaction, and user trust, along with general feedback on our final design.

B.2.1 Bayesian Knowledge Tracing

Instructions: Answer the following questions from memory and to the best of your ability (remember, we're evaluating the explainable, not you!). The questions are intended to vary in difficulty and length. Some may ask about similar concepts.

1. I have closed the BKT explainable.*

   ○ Yes

2. **Warmup:** List the parameters used in BKT.*

3. How well do you know your BKT parameters?*

   - Define **P(slip):**
   - Define **P(guess):**
   - What is the typical *upper* bound for **P(guess)**?
   - What is the typical *upper* bound for **P(slip)**?

4. **All the questions on this page pertain to the following scenario:** Sandy has a geoscience test coming up, but they have missed many days of classes recently due to a fever. The test consists of free-response questions and Sandy is given immediate feedback regarding the correctness of their answers.

   (a) Identify a specific piece of evidence from the scenario above that is related to **P(transit)**. Feel free to directly copy and paste text from the scenario itself.*

   (b) Based on the evidence identified in (a), the magnitude of Sandy's **P(transit)** is likely to be...*

   ○ Low ○ High
   ○ Somewhere in the middle

   (c) Based on your assessment of **P(transit)** in (b), what would be a reasonable value for this probability (between 0 and 1)?*

5. **All the questions on this page pertain to the following scenario:** Rory is about to take a math test. Assume that prior to this math test, Rory starts with a **P(init)** of 0.5. The test consists of 10 free-response questions. Now assume that Rory gets question one correct.

   (a) What is Rory's initial **P(init)** value at the beginning of the scenario?*

   (b) Identify the most important piece of evidence from the scenario above that is related to **P(init)** (other than the initial value of this parameter). Feel free to directly copy and paste text from the scenario itself.*

   (c) Based on the evidence identified in (b), do you think Rory's **P(init)** is likely to change?*

   ○ Yes ○ No

   Please explain your answer to (c).

(d) How would Rory's P(init) change (increase or decrease)?

○ Increase ○ Decrease

Please explain your answer to (d).

(e) Based on your assessment of **P(init)** above, what would be a reasonable value for this probability after Rory answers the first question of the test?*

6. **All the questions on this page pertain to the following scenario:** Darby likes to read ahead in their math class until they are sure they have learned the material. One day, Darby is asked to do a worksheet on material they have already gone through.

(a) Identify a specific piece of evidence from the scenario above that is related to **P(init)**. Feel free to directly copy and paste text from the scenario itself.*

(b) Based on the evidence identified in (a), the magnitude of Darby's **P(init)** is likely to be...*

○ Low ○ High
○ Somewhere in the middle

(c) Based on your assessment of Darby's **P(init)** in (b), is 0.4 a reasonable value for P(init) in this situation?*

○ Yes ○ No

Please explain your answer to (c).

(d) Based on the situation described in this scenario (i.e., completing a math worksheet), what would be a reasonable threshold for mastery (between 0 and 1)? *

Please explain your answer to (d).

(e) Now contrast this with a situation where BKT is being used to decide which applicants should be accepted into medical school. What would be a reasonable threshold for mastery in this case (between 0 and 1)?*

Please explain your answer to (e).

7. **This question pertains to the following scenario:** Marc has just returned to school after a summer vacation. Their teacher gives them a handout with review problems about fractions. Marc got a very good grade on the test last year but cannot remember how to add fractions for their life. On this previous test, Marc got a **P(init)** of 0.97.

(a) If BKT is used to assess Marc's performance on these new review problems, why might the algorithm's estimate of mastery not be accurate in this case?*

8. **All the questions on this page pertain to the following scenario:** Burt and Ernie both take the same test. Incredibly, they not only get the same score, but their answers are identical. Burt and Ernie took the test at the same time, in separate locations. No cheating occurred. However, while Burt took 17 minutes to take the test, Ernie took 4 hours. After the test, BKT gave Burt a **P(init)** of 0.99.

   (a) Identify a specific piece of evidence from the scenario above that can be used to differentiate Burt and Ernie's test performance. Feel free to directly copy and paste text from the scenario itself.*

   (b) How does the evidence you identified in (a) pertain to learning/mastery?*

   (c) Can BKT account for the evidence you identified in (a)? That is, can this irregular information be fit into the standard 4 BKT parameters?*

   ○ Yes ○ No

   Please explain your answer to (c).

   (d) BKT gives Ernie a **P(init)** of 0.99 too. Based on your answers to (b) and (c), is this behavior exhibited by BKT is intuitive/sensible?*

   ○ Yes ○ No

   Please explain your answer to (d).

   (e) How could you improve BKT's behavior to make it more intuitive and sensible in this case?*

9. **This question pertains to the following scenario:** Pinto is taking a strange multiple-choice exam where **P(guess)** is 0.67. They also drank too much coffee and find it difficult to fill in the bubbles correctly on the answer sheet (i.e. They fill in "C" when they want to select "A"), so their **P(slip)** is 0.5.

   (a) Consider the given values of **P(guess)** and **P(slip)** and explain how this may contribute to unexpected model behavior.*

10. **Honor Code:** I affirm that I completed the post-test questions above on my own, without looking at the BKT explainable or any other unauthorized resources.*

    ○ Yes ○ No

B.2.2 Decision Scenarios

Instructions: You will now be presented with four scenarios where an AI or human decision-maker is used to assess a hypothetical student's performance. Each scenario will begin with some general context and be followed by a more specific situation. Please read through each scenario and then answer the questions that follow.

Note: All the questions in this section were added for our second study. The order of scenarios was randomized for each participant.

1. **General context:** At an elementary school, first graders complete a math worksheet each week. An AI algorithm assesses their performance on these worksheets to decide whether they have mastered the material.

   Specific scenario: Mars is a first grader at the elementary school. The AI algorithm evaluates their performance on this week's math worksheet.

   - How fair or unfair is it for Mars that the AI algorithm evaluates their performance on their math worksheets?*

     ○ Very unfair ○ Unfair ○ Neither fair nor unfair ○ Fair ○ Very fair

     *Feel free to explain/elaborate on your answer above concerning the **fairness** of this decision.*

   - How much do you trust that the AI algorithm makes good-quality assessments of Mars' performance on their math worksheets?*

     ○ Extreme distrust ○ More distrust than trust ○ Neither trust nor distrust ○ More trust ○ Extreme trust

     *Feel free to explain/elaborate on your answer above concerning your **trust** in this decision.*

2. **General context:** At an elementary school, second graders complete a spelling worksheet each week. Their teacher assesses their performance on these worksheets to decide whether they have mastered the material.

   Specific scenario: Opal is a second grader at the elementary school. Their teacher evaluates their performance on this week's spelling worksheet.

   - How fair or unfair is it for Opal that their teacher evaluates their performance on their spelling worksheets?*

     ○ Very unfair ○ Unfair ○ Neither fair nor unfair ○ Fair ○ Very fair

     *Feel free to explain/elaborate on your answer above concerning the **fairness** of this decision.*

   - How much do you trust that the teacher makes good-quality assessments of Opal's performance on their spelling worksheets?*

     ○ Extreme distrust ○ More distrust than trust ○ Neither trust nor distrust ○ More trust ○ Extreme trust

*Feel free to explain/elaborate on your answer above concerning your **trust** in this decision.*

3. **General context:** 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.

   Specific scenario: Clay applies to the medical school. The AI algorithm evaluates their performance on the entrance exam.

   - How fair or unfair is it for Clay that the AI algorithm evaluates their performance on their medical school entrance exam?*

     ○ Very unfair ○ Unfair ○ Neither fair nor unfair ○ Fair ○ Very fair

     *Feel free to explain/elaborate on your answer above concerning the **fairness** of this decision.*

   - How much do you trust that the AI algorithm makes good-quality assessments of Clay's performance on their medical school entrance exam?*

     ○ Extreme distrust ○ More distrust than trust ○ Neither trust nor distrust ○ More trust ○ Extreme trust

     *Feel free to explain/elaborate on your answer above concerning your **trust** in this decision.*

4. **General context:** At a law school, first-year applicants must complete an entrance exam. To be recommended for admission, applicants must score highly on this exam. A human grader assesses their performance on the entrance exam.

   Specific scenario: Emerson applies to the law school. The human grader evaluates their performance on the entrance exam.

   - How fair or unfair is it for Emerson that the human grader evaluates their performance on their law school entrance exam?*

     ○ Very unfair ○ Unfair ○ Neither fair nor unfair ○ Fair ○ Very fair

     *Feel free to explain/elaborate on your answer above concerning the **fairness** of this decision.*

   - How much do you trust that the human grader makes good-quality assessments of Emerson's performance on their law school entrance exam?*

○ Extreme distrust	○ More distrust than trust	○ Neither trust nor distrust	○ More trust	○ Extreme trust

*Feel free to explain/elaborate on your answer above concerning your **trust** in this decision.*

B.2.3 Explainable Evaluation

1. Based on the BKT explainable you just completed, please rate how strongly you agree/disagree with the following statements.*

Explainable Design	Strongly disagree				Strongly agree
I thought the explainable was easy to follow.	○	○	○	○	○
I found the various sections in this explainable were well integrated.	○	○	○	○	○
I found this explainable engaging and interesting.	○	○	○	○	○
I enjoyed completing this explainable.	○	○	○	○	○

- Feel free to explain/elaborate on your answers here. We welcome any other comments about our **Explainable Design** as well.

Explainable Effectiveness	Strongly disagree				Strongly agree
I thought this explainable provided an effective explanation of BKT.	○	○	○	○	○
I could generally explain how BKT works to another person.	○	○	○	○	○
I feel comfortable using BKT systems after completing this explainable.	○	○	○	○	○
I am interested in learning more about ASL.	○	○	○	○	○

- Feel free to explain/elaborate on your answers here. We welcome any other comments about our **Explainable Effectiveness** as well.

Assessing BKT	Strongly disagree				Strongly agree
BKT provides accurate estimates of skill mastery.	○	○	○	○	○
BKT provides reliable, unbiased estimates of skill mastery.	○	○	○	○	○
BKT systems are helpful to students/teachers and reflect their best interests.	○	○	○	○	○
BKT's estimates of skill mastery are logical and easy to understand.	○	○	○	○	○
I would use BKT to assess skill mastery in different settings.	○	○	○	○	○
In general, I trust BKT's estimates of skill mastery.	○	○	○	○	○

- Feel free to explain/elaborate on your answers here. We welcome any other comments about **BKT** as well.

General Attitudes toward AI	Strongly disagree				Strongly agree
I believe AI does more good than harm.	○	○	○	○	○
I support the increasing prevalence of AI in society.	○	○	○	○	○
I am excited about the future of AI.	○	○	○	○	○
It is important to educate the public on how AI systems work.	○	○	○	○	○
It is important to educate the public on the potential harms and limitations of AI systems.	○	○	○	○	○

- Feel free to explain/elaborate on your answers here. We welcome any other comments about **AI systems/algorithms** as well.

2. **Last Question:** Any other comments/feedback for us regarding this study?

Appendix C

Mappings to BKT Knowledge Components

Here, I will provide the mappings from our BKT explainable activities and post-test questions to illustrate how the CTA-derived KCs were targetted throughout my study. KC numbers correspond to those used in Section 3.2.

C.1 Explainable KC Mappings

The mappings in this section were completed by Mira Sneirson '22. Question titles correspond to activities in my BKT explainable, described in Sections 4.1 and 4.2. The full explainable is also available here: https://catherinesyeh.github.io/bkt-asl/. Corresponding figures in this thesis will also be mentioned, if relevant.

Module	Question	Knowledge Area	KCs
Intro to P(guess) *Figure 4.3*	Reading "'BKT"	Identifying Priors	1, 2
	Guess estimation		1, 2, 4
Intro to P(init)	Reading "Cat"	Identifying Priors	1, 2
		Evaluating P(init)	1
	What happened?	Identifying Priors	2
		Evaluating P(init)	1
	How did P(init) change?	Identifying Priors	1
		Identifying Changed Parameters	1, 2, 3, 4

Module	Question	Knowledge Area	KCs
Intro to P(transit) ***Figure 4.7***	Signing fast and slow Why is it more challenging? Learning difficulty	Identifying Priors	1, 2
	What impacts P(transit)?		2
Intro to P(slip)	Similarity of T & S Other similar letters	Identifying Priors	1, 2
Identifying Parameters ***Figure 4.5***	Which parameter is which?	Identifying Priors Evaluating P(init)	1, 2 1
Working to Mastery ***Figure 4.8***	Achieve mastery!	Identifying Priors Identifying Changed Parameters Evaluating P(init)	1 1, 2, 4 1
Losing Mastery ***Figure 4.9***	How does P(init) change over time?	Identifying Priors Evaluating P(init) Limitations of BKT	3, 5 2 1, 2
	When does P(init) start to decrease?	Identifying Priors Evaluating P(init) Limitations of BKT	3 2 1, 2
	Summer vacation	Identifying Priors Evaluating P(init) Limitations of BKT	3 2, 3, 4, 5 2
Role of Speed	Mac & Cheese scenario	Identifying Priors Evaluating P(init)	3 1, 2, 3, 4, 5
	Real-life reasoning	Identifying Priors Evaluating P(init) Limitations of BKT	3 4 2, 3

Module	Question	Knowledge Area	KCs
Incorrect Answers	See what happens when you're wrong	Identifying Changed Parameters	1, 2, 4
		Evaluating P(init)	1
	Why would an increase make sense?	Evaluating P(init)	1, 3
		Limitations of BKT	3
	What should happen?	Evaluating P(init)	4, 5
		Limitations of BKT	3
Model Degeneracy ***Figure 4.6***	When P(slip) and/or P(guess) > 0.5	Identifying Priors	1, 3
		Identifying Changed Parameters	3, 4
		Evaluating P(init)	4
		Limitations of BKT	4
	Apply that logic to P(guess)	Identifying Priors	1, 2, 3
		Evaluating P(init)	4
		Limitations of BKT	4

C.2 Post-Test KC Mappings

The mappings in this section (found on the next page) correspond to the post-test questions assessing user mental models of BKT after completing my explainable, as discussed in Section 4.3.1. All question numbers should correspond to those used in Section B.2.1. Questions 1 and 10 are omitted since they pertain to our participant honor code rather than assessing understanding of BKT.

Question	Knowledge Area	Knowledge Component
2, 3	Identifying Priors	1
4a		2
4b		4
4c		5
5a	Identifying Changed Parameters	1
5b		2
5c		3
5d		4
5e		5
6a	Evaluating P(init)	1
6b		2
6c		3
6d, 6e		4
7		5
8a	Limitations of BKT	1
8b, 8c		2
8d, 8e		3
9		4

Bibliography

[1] ACM FAccT. Acm conference on fairness, accountability, and transparency. `https://facctconference.org/`. Accessed: 2021-10-14.

[2] Adar, E., and Lee, E. Communicative visualizations as a learning problem. *IEEE Transactions on Visualization and Computer Graphics 27*, 2 (2020), 946–956.

[3] Agarwal, D., Babel, N., and Baker, R. S. Contextual derivation of stable bkt parameters for analysing content efficacy. In *EDM* (2018).

[4] Aguinis, H., Villamor, I., and Ramani, R. S. Mturk research: Review and recommendations. *Journal of Management 47*, 4 (2021), 823–837.

[5] Allen, I. E., and Seaman, C. A. Likert scales and data analyses. *Quality progress 40*, 7 (2007), 64–65.

[6] Anderson, A., Dodge, J., Sadarangani, A., Juozapaitis, Z., Newman, E., Irvine, J., Chattopadhyay, S., Olson, M., Fern, A., and Burnett, M. Mental models of mere mortals with explanations of reinforcement learning. *ACM Transactions on Interactive Intelligent Systems (TiiS) 10*, 2 (2020), 1–37.

[7] Anik, A. I., and Bunt, A. Data-centric explanations: Explaining training data of machine learning systems to promote transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–13.

[8] Ardito, C., De Marsico, M., Lanzilotti, R., Levialdi, S., Roselli, T., Rossano, V., and Tersigni, M. Usability of e-learning tools. In *Proceedings of the working conference on Advanced visual interfaces* (2004), pp. 80–84.

[9] Arya, V., Bellamy, R. K., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

[10] Atzmüller, C., and Steiner, P. M. Experimental vignette studies in survey research. *Methodology* (2010).

[11] Badrinath, A., Wang, F., and Pardos, Z. pybkt: An accessible python library of bayesian knowledge tracing models. *arXiv preprint arXiv:2105.00385* (2021).

[12] Bandura, A., et al. Guide for constructing self-efficacy scales. *Self-efficacy beliefs of adolescents 5*, 1 (2006), 307–337.

[13] Bangor, A., Kortum, P. T., and Miller, J. T. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction 24*, 6 (2008), 574–594.

[14] Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–16.

[15] Barnett, S., McKee, M., Smith, S., and Pearson, T. Deaf sign language users, health inequities, and public health. *Opportunity for social justice* (2011), 8.

[16] Barron, K. E., and Harackiewicz, J. M. Achievement goals and optimal motivation: testing multiple goal models. *Journal of personality and social psychology 80*, 5 (2001), 706.

[17] Bassen, J., Howley, I., Fast, E., Mitchell, J., and Thille, C. Oars: exploring instructor analytics for online learning. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (2018), ACM, p. 55.

[18] Berkovsky, S., Taib, R., and Conway, D. How to recommend? user trust factors in movie recommender systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (2017), pp. 287–300.

[19] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M., and Eckersley, P. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (2020), pp. 648–657.

[20] Bier, N., Lip, S., Strader, R., Thille, C., and Zimmaro, D. An approach to knowledge component/skill modeling in online courses. *Open Learning* (2014), 1–14.

[21] Bjork, R. A. Memory and metamemory considerations in the training of human beings. *Metacognition: Knowing about knowing 185*, 7.2 (1994).

[22] Buçinca, Z., Malaya, M. B., and Gajos, K. Z. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction 5*, CSCW1 (2021), 1–21.

[23] Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G. S., Stumpe, M. C., et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems* (2019), pp. 1–14.

[24] Clark, R. E., Feldon, D. F., van Merriënboer, J. J. G., Yates, K., and Early, S. Cognitive task analysis. In *Handbook of research on educational communications and technology: a project of the association for educational communications and technology.* Routledge, New York, NY USA, 2008, p. 577–593.

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

[26] Council, N. R., et al. How experts differ from novices. In *How people learn: Brain, mind, experience, and school: Expanded edition.* National Academies Press, 2000, pp. 31–50.

[27] d Baker, R. S., Corbett, A. T., and Aleven, V. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *Intelligent Tutoring Systems* (2008), Springer, pp. 406–415.

[28] Davies, I., Green, P., Rosemann, M., Indulska, M., and Gallo, S. How do practitioners use conceptual modeling in practice? *Data & Knowledge Engineering 58*, 3 (2006), 358–380.

[29] Deonovic, B., Yudelson, M., Bolsinova, M., Attali, M., and Maris, G. Learning meets assessment. *Behaviormetrika 45*, 2 (2018), 457–474.

[30] Do, C. B., and Batzoglou, S. What is the expectation maximization algorithm? *Nature biotechnology 26*, 8 (2008), 897–899.

[31] Doorley, S., Holcomb, S., Klebahn, P., Segovia, K., and Utley, J. Design thinking bootleg, 2018.

[32] Doroudi, S., and Brunskill, E. The misidentified identifiability problem of bayesian knowledge tracing. *International Educational Data Mining Society* (2017).

[33] Doroudi, S., and Brunskill, E. Fairer but not fair enough on the equitability of knowledge tracing. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (2019), pp. 335–339.

[34] Dos Santos, D. P., Giese, D., Brodehl, S., Chon, S., Staab, W., Kleinert, R., Maintz, D., and Baessler, B. Medical students' attitude towards artificial intelligence: a multicentre survey. *European radiology 29*, 4 (2019), 1640–1646.

[35] Doshi-Velez, F., and Kim, B. A roadmap for a rigorous science of interpretability. *arXiv preprint arXiv:1702.08608 2* (2017).

[36] Du, M., Liu, N., and Hu, X. Techniques for interpretable machine learning. *Communications of the ACM 63*, 1 (2019), 68–77.

[37] Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., and Giro-i Nieto, X. How2sign: a large-scale multimodal dataset for

continuous american sign language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 2735–2744.

[38] Dweck, C. S. *Mindset: The new psychology of success*. Random house, 2006.

[39] Edelson, D. C., Gordin, D. N., and Pea, R. D. Addressing the challenges of inquiry-based learning through technology and curriculum design. *Journal of the learning sciences 8*, 3-4 (1999), 391–450.

[40] Egelman, S., and Peer, E. Scaling the security wall: Developing a security behavior intentions scale (sebis). In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (2015), pp. 2873–2882.

[41] Evans, R. J., and Richardson, T. S. Maximum likelihood fitting of acyclic directed mixed graphs to binary data. *arXiv preprint arXiv:1203.3479* (2012).

[42] Fast, L. A., Lewis, J. L., Bryant, M. J., Bocian, K. A., Cardullo, R. A., Rettig, M., and Hammond, K. A. Does math self-efficacy mediate the effect of the perceived classroom environment on standardized math test performance? *Journal of educational psychology 102*, 3 (2010), 729.

[43] Fu, F.-L., Su, R.-C., and Yu, S.-C. Egameflow: A scale to measure learners' enjoyment of e-learning games. *Computers & Education 52*, 1 (2009), 101–112.

[44] Gade, K., Geyik, S. C., Kenthapadi, K., Mithal, V., and Taly, A. Explainable ai in industry. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (2019), pp. 3203–3204.

[45] González-Brenes, J., Huang, Y., and Brusilovsky, P. General features in knowledge tracing to model multiple subskills, temporal item response theory, and expert knowledge. In *The 7th international conference on educational data mining* (2014), University of Pittsburgh, pp. 84–91.

[46] Hadar, L., Danziger, S., and Hertwig, R. The attraction effect in experience-based decisions. *Journal of Behavioral Decision Making 31*, 3 (2018), 461–468.

[47] Hake, R. R. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American journal of Physics 66*, 1 (1998), 64–74.

[48] Hauser, D. J., and Schwarz, N. Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior research methods 48*, 1 (2016), 400–407.

[49] Hawkins, W. J., Heffernan, N. T., and Baker, R. S. Learning bayesian knowledge tracing parameters with a knowledge heuristic and empirical probabilities. In *International Conference on Intelligent Tutoring Systems* (2014), Springer, pp. 150–155.

[50] Hays, M. J., Kornell, N., and Bjork, R. A. When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition 39*, 1 (2013), 290.

[51] Heerink, M., Kröse, B., Evers, V., and Wielinga, B. The influence of social presence on acceptance of a companion robot by older people. *Journal of Physical Agents 2*, 2 (2008), 33–40.

[52] Hogan, K. E., and Pressley, M. E. *Scaffolding student learning: Instructional approaches and issues.* Brookline Books, 1997.

[53] Holstein, K., McLaren, B. M., and Aleven, V. Student learning benefits of a mixed-reality teacher awareness tool in ai-enhanced classrooms. In *International Conference on Artificial Intelligence in Education* (2018), Springer, pp. 154–168.

[54] Huff, C., and Tingley, D. "who are these people?" evaluating the demographic characteristics and political preferences of mturk survey respondents. *Research & Politics 2*, 3 (2015), 2053168015604648.

[55] Hutchison, M. A., Follman, D. K., Sumpter, M., and Bodner, G. M. Factors influencing the self-efficacy beliefs of first-year engineering students. *Journal of Engineering Education 95*, 1 (2006), 39–47.

[56] Inan, H. Z., and Inan, T. 3 h s education: Examining hands-on, heads-on and hearts-on early childhood science education. *International Journal of Science Education 37*, 12 (2015), 1974–1991.

[57] Jain, P., and Kar, P. Non-convex optimization for machine learning. *arXiv preprint arXiv:1712.07897* (2017).

[58] Kizilcec, R. F. How much information? effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), pp. 2390–2395.

[59] Koedinger, K. R., Corbett, A. T., and Perfetti, C. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science 36*, 5 (2012), 757–798.

[60] Koedinger, K. R., Kim, J., Jia, J. Z., McLaughlin, E. A., and Bier, N. L. Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In *Proceedings of the second (2015) ACM conference on learning@ scale* (New York, NY USA, 2015), Association for Computing Machinery, pp. 111–120.

[61] Lakkaraju, H., and Bastani, O. "how do i fool you": Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020), AIES '20, Association for Computing Machinery, p. 79–85.

[62] Lawson, B. *How designers think: The design process demystified*. Routledge, 2006.

[63] Lee, M. K. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society 5*, 1 (2018), 2053951718756684.

[64] Lee, Y., Kozar, K. A., and Larsen, K. R. The technology acceptance model: Past, present, and future. *Communications of the Association for information systems 12*, 1 (2003), 50.

[65] Lehrer, K. *Theory of knowledge*. Routledge, 2018.

[66] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy 23*, 1 (2021), 18.

[67] Lipton, Z. C. The mythos of model interpretability. *Queue 16*, 3 (2018), 31–57.

[68] Lovett, M. C. Cognitive task analysis in service of intelligent tutoring system design: A case study in statistics. In *International Conference on Intelligent Tutoring Systems* (1998), Springer, pp. 234–243.

[69] Lu, J., Lee, D. D., Kim, T. W., and Danks, D. Good explanation for algorithmic transparency. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020), AIES '20, Association for Computing Machinery, p. 93.

[70] Mahatody, T., Sagar, M., and Kolski, C. State of the art on the cognitive walkthrough method, its variants and evolutions. *Intl. Journal of Human–Computer Interaction 26*, 8 (2010), 741–785.

[71] McGovern, A., and Fager, J. Creating significant learning experiences in introductory artificial intelligence. In *Proceedings of the 38th SIGCSE technical symposium on Computer science education* (2007), pp. 39–43.

[72] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (2019), pp. 220–229.

[73] Mohseni, S., Zarei, N., and Ragan, E. D. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS) 11*, 3-4 (2021), 1–45.

[74] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences 116*, 44 (2019), 22071–22080.

[75] National Institute of Deafness and Other Communication Disorders. American sign language. `https://www.nidcd.nih.gov/health/american-sign-language`, 2019. Accessed: 2022-10-1.

[76] Nielsen, J. *Usability engineering*. Morgan Kaufmann, 1994.

[77] Nielsen, J., and Molich, R. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (1990), pp. 249–256.

[78] Nisbett, R. E., Caputo, C., Legant, P., and Marecek, J. Behavior as seen by the actor and as seen by the observer. *Journal of personality and Social Psychology 27*, 2 (1973), 154.

[79] Oppenheimer, D. M., Diemand-Yauman, C., and Vaughan, E. B. Fortune favors the bold (and the italicized): Effects of disfluency on educational outcomes. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (2010).

[80] Pardos, Z., Bergner, Y., Seaton, D., and Pritchard, D. Adapting bayesian knowledge tracing to a massive open online course in edx. In *Educational Data Mining 2013* (2013), Citeseer.

[81] Pardos, Z. A., and Heffernan, N. T. Modeling individualization in a bayesian networks implementation of knowledge tracing. In *International conference on user modeling, adaptation, and personalization* (2010), Springer, pp. 255–266.

[82] Pardos, Z. A., and Heffernan, N. T. Kt-idem: Introducing item difficulty to the knowledge tracing model. In *International conference on user modeling, adaptation, and personalization* (2011), Springer, pp. 243–254.

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

[84] Penney, S., Dodge, J., Anderson, A., Hilderbrand, C., Simpson, L., and Burnett, M. The shoutcasters, the game enthusiasts, and the ai: Foraging for explanations of real-time strategy players. *ACM Transactions on Interactive Intelligent Systems (TiiS) 11*, 1 (2021), 1–46.

[85] Phan, M. H., Keebler, J. R., and Chaparro, B. S. The development and validation of the game user experience satisfaction scale (guess). *Human factors 58*, 8 (2016), 1217–1247.

[86] Pirolli, P., and Card, S. Information foraging. *Psychological review 106*, 4 (1999), 643.

[87] Plass, J. L., Moreno, R., and Brünken, R. *Cognitive load theory*. Cambridge university press, 2010.

[88] Polson, P. G., Lewis, C., Rieman, J., and Wharton, C. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of man-machine studies 36*, 5 (1992), 741–773.

[89] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. Manipulating and measuring model interpretability. In *Proceedings*

of the 2021 CHI conference on human factors in computing systems (New York, NY USA, 2021), Association for Computing Machinery, pp. 1–52.

[90] Qiu, Y., Qi, Y., Lu, H., Pardos, Z. A., and Heffernan, N. T. Does time matter? modeling the effect of time with bayesian knowledge tracing. In *EDM* (2011), pp. 139–148.

[91] Rettig, M. Prototyping for tiny fingers. *Communications of the ACM 37*, 4 (1994), 21–27.

[92] Rieber, L. P. The effect of logo on increasing systematic and procedural thinking according to piaget's theory of intellectual development and on its ability to teach geometric concepts to young children. *Journal of Educational Computing Research* (1983), 249–260.

[93] Rosen, Y., Rushkin, I., Rubin, R., Munson, L., Ang, A., Weber, G., Lopez, G., and Tingley, D. The effects of adaptive learning in a massive open online course on learners' skill development. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (2018), pp. 1–8.

[94] Saha, D., Schumann, C., McElfresh, D. C., Dickerson, J. P., Mazurek, M. L., and Tschantz, M. C. Human comprehension of fairness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020), AIES '20, Association for Computing Machinery, p. 152.

[95] Sani, N., Lee, J., Nabi, R., and Shpitser, I. A semiparametric approach to interpretable machine learning. *arXiv preprint arXiv:2006.04732* (2020).

[96] Snyder, C. *Paper prototyping: The fast and easy way to design and refine user interfaces.* Morgan Kaufmann, 2003.

[97] Stallings, J. Allocated academic learning time revisited, or beyond time on task. *Educational researcher 9*, 11 (1980), 11–16.

[98] Šumak, B., Heričko, M., and Pušnik, M. A meta-analysis of e-learning technology acceptance: The role of user types and e-learning technology types. *Computers in human behavior 27*, 6 (2011), 2067–2077.

[99] Suresh, H., Gomez, S. R., Nam, K. K., and Satyanarayan, A. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–16.

[100] Turner, C. W., Lewis, J. R., and Nielsen, J. Determining usability test sample size. *International encyclopedia of ergonomics and human factors 3*, 2 (2006), 3084–3088.

[101] van Berkel, N., Goncalves, J., Russo, D., Hosio, S., and Skov, M. B. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–13.

[102] Van Joolingen, W. R., and De Jong, T. Supporting hypothesis generation by learners exploring an interactive computer simulation. *Instructional Science 20*, 5 (1991), 389–404.

[103] Van Someren, M., Barnard, Y., and Sandberg, J. *The think aloud method: a practical approach to modelling cognitive*. Citeseer, 1994.

[104] VanLehn, K., Jones, R. M., and Chi, M. T. A model of the self-explanation effect. *The journal of the learning sciences 2*, 1 (1992), 1–59.

[105] Visxai.io. Visualization for ai explainability. `https://visxai.io/`. Accessed: 2021-10-14.

[106] Wang, C., Kim, D.-H., Bai, R., and Hu, J. Psychometric properties of a self-efficacy scale for english language learners in china. *System 44* (2014), 24–33.

[107] Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), CHI '19, Association for Computing Machinery, p. 1–15.

[108] Watkins, R., Meiers, M. W., and Visser, Y. Cognitive task analysis. In *A guide to assessing needs: Essential tools for collecting information, making decisions, and achieving development results*. World Bank Publications, 2012, pp. 156–164.

[109] Wiggins, G., Wiggins, G. P., and McTighe, J. *Understanding by design*. Ascd, 2005.

[110] Williamson, K., and Kizilcec, R. F. Effects of algorithmic transparency in bayesian knowledge tracing on trust and perceived accuracy. *International Educational Data Mining Society* (2021).

[111] Yeh, C., and Howley, I. Cognitive task analysis for empirical post-hoc ai explanations. In *Grace Hopper Celebration of Women in Computing* (2021), ACM Student Research Competition.

[112] Yudelson, M. V., Koedinger, K. R., and Gordon, G. J. Individualized bayesian knowledge tracing models. In *International conference on artificial intelligence in education* (2013), Springer, pp. 171–180.

[113] Zhang, B., and Dafoe, A. Artificial intelligence: American attitudes and trends. *Available at SSRN 3312874* (2019).

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