Meaningful Models: Unlocking Insights Through Model Interpretations

Napol Rachatasumrit

CMU-HCII-25-103
June 2025

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Dr. Kenneth Koedinger
Dr. Paulo Carvalho
Dr. Adam Sales
Dr. Kenneth Holstein

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Keywords: Knowledge Tracing, Interpretable Machine Learning, Explainable AI, Student Modeling, Simulated Learners

Abstract

The conventional wisdom in Educational Data Mining (EDM) suggests that a superior model fits the data better. However, this perspective overlooks a critical aspect: the value of machine learning models lies not merely in their predictive power, but fundamentally comes from their use. Models that prioritize prediction accuracy often fail to provide scientifically or practically meaningful interpretations. Meaningful interpretations are crucial for scientific insight and often yield practical applications, especially from the human-centered perspective. For example, a popular knowledge tracing model using deep learning has been demonstrated to have a superior predictive power of student performance; however, its parameters do not have an association with any latent constructs, so there have been no scientific insights or practical applications resulting from it. In contrast, a logistic regression model often underperforms its deep learning counterparts in prediction accuracy, but its parameter estimates have meaningful interpretations (e.g., the slope illustrates the rate of learning of knowledge components) that lead to new scientific insights (e.g. improved cognitive models discovery) and results in useful practical applications (e.g. an intelligent tutoring system redesign).

In this thesis, I argue for a claim that meaningful interpretations are what we need rather than post-hoc explanations or uninterpreted interpretable models, especially in the context of EDM. I explore a concept of "meaningful models" as inherently interpretable models whose parameters and outputs are not only transparent but actively interpreted. Moreover, their interpretations lead to useful and actionable insights for stakeholders. I illustrate the benefits of meaningful models through examples where existing mechanisms or models are insufficient to produce meaningful interpretations and demonstrating how enhancements can yield scientifically or practically valuable insights. For example, Performance Factor Analysis (PFA) has been demonstrated to outperform its base model, but we show that PFA parameters are confounded, which resulted in ambiguous interpretations. We then proposed improved models that not only de-confound the parameters but also presented meaningful interpretations that lead to insights on the associated knowledge component model and suggested instructional improvement. Overall, this thesis highlights the essential role of meaningful models in EDM, emphasizing that only through meaningful interpretations can models effectively drive practical improvements in educational practices and advance scientific understanding.

Acknowledgements

First and foremost, I would like to thank my parents and my brother for their love and support, and everything they have done for me through these years. They always support my decision with love and encouragement. To my mom for always being there for me with her positivity through good and bad times. To my dad for being my role model who has taught me the importance of learning and motivated me to always try my best. To my brother for always quietly supporting me to be able to follow my passions. None of this would be possible without you guys.

I would also like to express my gratitude to faculty members, staff, and friends who I have encountered throughout my time in the PhD program. I appreciate all the help and good times we have shared. I would especially like to thank Danny for the many great conversations, and I am fortunate to get to work with you as my labmate. I would like to thank my advisors, Ken and Paulo, for so many great years of teaching and inspiring me to be a better researcher. I have grown so much from working with both of you through these years.

I would like to thank God for being my savior and guiding me through all the difficulties. Finally, I would like to give special thanks to my wife, Belle, for her continuous support and understanding through these years, including so many sleepless nights of LDR. Thank you for putting up with my selfishness and stubbornness to finish a PhD, and I really appreciate everything you have done for me!

Table of Contents

EDM	Educational Data Mining
DKT	Deep Knowledge Tracing
PFA	Performance Factor Analysis
BKT	Bayesian Knowledge Tracing
AFM	Additive Factor Model
ITS	Intelligent Tutoring System
RMSE	Root Mean Squared Error
BIC	Bayesian Information Criterion
AUC	Area Under the Receiver Operating Characteristic Curve
RNN	Recurrent Neural Network
KC	Knowledge Component
IRT	Item Response Theory

Chapter 1
Introduction

Educational Data Mining (EDM) is a crucial field in learning sciences that leverages data analysis to enhance educational outcomes and personalize learning experiences. By analyzing large amounts of educational data, EDM researchers can discover patterns and insights that lead to improvements in pedagogy design, curriculum development, and student intervention. One prominent example of EDM is student modeling with knowledge tracing — models that estimate students' mastery of specific skills over time, which has been widely used in Intelligent Tutoring Systems (ITS) to adaptively assess students' knowledge states.

The recent trend in EDM, and in data mining more generally, suggests that a superior model fits the data better [71]. In other words, a model that performs better on fit statistics, such as root mean squared error (RMSE) [9], bayesian information criterion (BIC) [23], or area under the receiver operating characteristic curve (AUC) [8], is usually considered a better model. However, this perspective overlooks a critical aspect: the value of machine learning models lies not merely in their predictive power, but fundamentally comes from their use. In other words, a model is not just a mathematical artifact; it is a tool, and its worth is determined by how it contributes to stakeholders' benefits. Models that prioritize prediction accuracy often fail to provide scientifically or practically meaningful interpretations, which are crucial for scientific insight and often yield practical applications. While accurate prediction could be useful, the focus on prediction accuracy alone can overshadow the importance of understanding and prioritizing the needs and benefits of stakeholders who will use these models [36].

My goal for this thesis is to reconsider interpretable machine learning through a human-centered approach. By emphasizing stakeholder utility, it is essential to move beyond post-hoc explanations of black-box models and focus instead on developing inherently interpretable models. These models should offer not only improved predictive accuracy but also parameter estimates that clearly explain their predictions. Specifically, parameters in such meaningful models should align with latent variables, thus providing valuable insights into the educational processes they represent. Nevertheless, it is insufficient to merely identify these parameters; meaningful interpretations derived from these models are necessary to genuinely deliver practical benefits to stakeholders.

Why are these meaningful interpretations important to stakeholders? These interpretations are crucial for scientific insight and are useful for practical applications,

especially from the human-computer interaction perspective. For example, Deep Knowledge Tracing (DKT) [57], a knowledge tracing model based on Recurrent Neural Network (RNN) [73], has been demonstrated to predict student performance better than traditional approaches based on logistic regression [1]; however, its parameters do not have an association with any latent constructs, so it lacks meaningful interpretation and consequently provides neither scientific insights nor practical applications [70]. In contrast, Additive Factor Model (AFM) [13], a knowledge tracing model based on logistic regression, often underperforms DKT in prediction accuracy, but its parameter estimates have meaningful interpretations (e.g., the slope illustrates the rate of learning of knowledge components) that lead to new scientific insights (e.g. improved cognitive models discovery) and results in useful practical applications (e.g. an intelligent tutoring system redesign) [41].

It is likely a misconception that complex black-box models are always superior in terms of predictive performance. In many cases, simpler, interpretable models can achieve comparable accuracy [23, 35, 56, 62, 84, 92], while still providing valuable insights into the learning mechanisms and pedagogy [33, 41]. For example, it has been shown that a logistic regression model, with the right set of features, was as good as DKT in predicting student performances on several datasets, while also preserving the meaningful interpretation of their parameter estimates [23, 44, 62]. Emphasizing the development and use of inherently interpretable models in EDM can lead to more effective and actionable educational interventions. More examples from my previous works are discussed further in Chapter 5.

In this thesis, I argue that post-hoc explanations and inherently interpretable models are not sufficient without being interpreted. Particularly in high-stakes domains such as EDM, the utility of these models is significantly limited without meaningful interpretation. To address this issue, I introduce the concept of "meaningful models," defined as models whose interpretations yield actionable insights or facilitate scientific discovery. To highlight the urgency of this issue , especially within the EDM community, this thesis includes a review of recent publications from leading AIED and EDM conferences, revealing that most proposed models emphasize predictive performance, and even those that claim interpretability often do not actively engage with their interpretations, thus significantly reducing their practical utility. Furthermore, this thesis provides multiple examples demonstrating how meaningful models and their interpretations can produce actionable insights, guide pedagogical strategies, and advance scientific understanding.

This document is organized as follows: Chapter 2 provides background information and discusses related work, beginning with an overview of literature concerning model explainability and interpretability, emphasizing their relevance within EDM contexts. It also reviews the historical development and existing research on knowledge tracing models and models of human learning. In Chapter 3, I present my central argument regarding the necessity for meaningful models, constructing a framework grounded in literature from

multiple related disciplines, including explainable and interpretable machine learning, human-computer interaction, cognitive science, and learning sciences. Chapter 4 critiques predictive performance metrics, illustrating their unreliability and potential for misleading conclusions within educational contexts. Through experiments using synthetic data, I demonstrate that knowledge tracing models selected based on superior fit statistics, such as the Bayesian Information Criterion (BIC), may perform poorly when implemented in practical applications, such as adaptive learning systems. Chapters 5 and 6 explore various methodologies for constructing and interpreting meaningful models, demonstrating their potential to produce actionable insights beneficial to stakeholders, including learning scientists and educators. Finally, the thesis concludes by summarizing these contributions and their implications.

Chapter 2
Background

2.1 Model Interpretability

The widespread use of black-box machine learning models in high-stakes decision-making areas, such as healthcare and criminal justice, has led to significant challenges and ethical concerns [87]. Similarly the field of EDM has prioritized prediction accuracy such that black-box models have been increasingly used [20, 25]. However, black-box models not only present challenges for applications in high-stakes domains, but also fail, by themselves, to provide useful insights, scientifically or practically.

Interpretability in machine learning encompasses multiple dimensions, reflecting diverse viewpoints on the importance and methods of making models understandable to humans. It involves both the transparency of a model's internal mechanisms and the clarity of its outputs, aiming to bridge the gap between complex algorithms and human comprehension. This multifaceted nature underscores the need for clear definitions and context-specific approaches to ensure that interpretability efforts align with the goals of various stakeholders.

To address the multifaceted nature of interpretability in machine learning, Murdoc et al. introduced The PDR framework offering a structured approach to evaluating interpretability in machine learning by emphasizing three key criteria: predictive accuracy, descriptive accuracy, and relevancy [51]. Predictive accuracy assesses how well a model generalizes to new data, ensuring its outputs are reliable. Descriptive accuracy measures how faithfully an interpretation reflects the model's internal mechanisms and learned relationships. Relevancy considers whether the interpretation provides meaningful and actionable insights for a specific human audience within a given domain. The framework also distinguishes between model-based interpretability methods, which are inherently transparent, and post-hoc methods that explain black-box models after training. By integrating these dimensions, the PDR framework aids practitioners in selecting and evaluating interpretability techniques that align with their specific application needs and audience requirements .

Lipton et al. also argues that interpretability is not a monolithic concept but encompasses various distinct ideas, leading to confusion and inconsistent claims in the literature [26]. The authors critically examine the ambiguous and multifaceted nature of interpretability in machine learning. They identify two primary notions: transparency, where

a model's internal mechanics are inherently understandable, and post-hoc explanations, which provide insights after model training. They also challenge the assumption that linear models are inherently interpretable, noting that factors like feature engineering can complicate their interpretability. Ultimately, the authors call for a more precise and context-dependent understanding of interpretability, urging researchers to clearly define what they mean by interpretability and to align their methods with the specific needs of the application domain.

The discourse around interpretability in machine learning has given rise to different schools of thought. One perspective advocates for post-hoc explanation methods, which attempt to elucidate the decisions of black-box models after they have been trained. These methods include techniques like SHAP values [42], LIME [22], and saliency maps, aiming to provide insights into model behavior. While both Murdoch et al. and Lipton explicitly distinguish between inherently interpretable models and post-hoc explanations, they acknowledge that post-hoc methods can mitigate these issues [32], this approach often perpetuates problematic practices. In response, Rudin et al. have proposed that the preferable strategy is to design models that are inherently interpretable by design [71]. This perspective underscores the fundamental difference between explaining black-box models and using inherently interpretable models, such that explanation is post hoc and does not lead to the understanding of the underlying mechanisms of the events or the nature of the data. Instead, meaningful models provide transparency and accountability, which are crucial in applications that directly impact stakeholders and could lead to useful insights.

The problem is that it is almost always easier to find an accurate-but-complex model than an accurate-yet-simple model. However, Semenova et al. pointed out that, given a predictive model, there is usually a large equivalence set of similarly accurate models known as the Rashomon set. This set includes some models that are highly parameterized and difficult to understand, while others are simpler and more interpretable [74]. Therefore, given an accurate black-box model, an inherently interpretable model is likely to exist but unlikely to be produced by deep learning.

Usually, a machine learning model would be considered interpretable when it is simple enough (e.g. smaller number of parameters) for humans to comprehend and understand the relationship between input features and output prediction. However, in the context of this thesis, I aim to expand on the definition of interpretable models to "meaningful models", such that the input features themselves need to be meaningful and represent some latent constructs. Moreover, the parameter estimates from meaningful models should provide insights that lead to the understanding of the underlying mechanisms or practical applications. For instance, consider a simple linear regression model predicting the probability of diabetes. If one of its features is a complex and arbitrary computation, such as weight multiplied by the number of siblings, the model may not be genuinely interpretable. Even though the model predicts an outcome, the inclusion of obscure or unrelated features can obscure its interpretability, making it challenging to

understand how and why certain predictions are made. Chapter 3 discusses my reasonings and arguments on the importance of meaningful models that are not only interpretable but are interpreted in ways that yield practical or scientific insight.

2.2 Knowledge Tracing and Models of Learning

The main objective of EDM is to improve educational systems by applying data mining techniques to educational data, such as student interactions with an ITS, to obtain useful insights, especially on the students' learning processes. These insights can then help refine teaching strategies and enhance student achievement. Knowledge tracing models are among the most popular models that have been explored in the field of EDM. These models take students' past performance on related problems associated with a set of knowledge components [34], as inputs and output predicted student performance on a particular problem or a student's mastery on a certain knowledge component.

Traditionally, there are two popular approaches to knowledge tracing models. Early attempts based on a Bayesian inference approach, which usually relied on simplifying the model assumptions (e.g. student's mastery is a binary state). Bayesian Knowledge Tracing (BKT) [17], which models student mastery as a latent variable in a simple Hidden Markov Model [7], has been widely used in the real-world ITSs and shown to be reasonably effective for mastery learning and problem selection [75]. Another popular approach to knowledge tracing models is a series of models based on logistic regression models, such as Additive Factor Model (AFM) and Performance Factor Analysis (PFA) [55]. In contrast to BKT, these models do not assume student mastery as a binary variable but use a parametric factor analysis approach to trace a student's knowledge based on a variety of factors, such as number of previous opportunities. Recently, with the rising popularity of neural networks, a large number of knowledge tracing models based on different deep learning techniques has been introduced. Deep Knowledge Tracing (DKT) is the pioneer of the deep learning based approach, which is based on a sequence model called Recurrent Neural Network (RNN). In the earlier works, DKT has been demonstrated to outperform the existing models, such as BKT and PFA, in many scenarios. However, recent work has further studied its pitfalls and showed that these deep learning models do not always outperform traditional models; model success depends on the nature of the dataset [23, 62, 84, 92].

In the context of interpretability, traditional models based on Bayesian inference and logistic regression usually have parameters that have meaningful interpretations, intentionally or not, due to the simplicity of the models and variables that are based on related latent constructs, such as a probability that a student makes a mistake when applying a known skill or a probability that a student guesses an answer correctly. However, deep learning based knowledge tracing models often forgo interpretability for potentially stronger predictive power due to the extremely large amount of parameters that

these models usually have. On a related note, the traditional evaluation methods for knowledge tracing models have focused on goodness-of-fit (e.g. AIC and BIC) and cross-validation. However, recent trends emphasize the use of metrics like AUC. This shift is driven by the increasing complexity and number of parameters in the deep learning based models, which have a strong negative impact on metrics like BIC. In my previous work, it is demonstrated that relying solely on AUC might not always accurately represent the quality of knowledge tracing models in the practical applications [62].

It could be argued that the primary goal of knowledge tracing is to predict student outcomes accurately. However, prior studies indicate that we can gain much more from parameter estimates, providing deeper insights into learning processes [33, 41]. If the objective of EDM is to enhance our understanding of learning, which leads to improved student outcomes, models that naively predict student's performance without offering interpretability that can result in useful insights could be considered inadequate. Despite the utility of such predictions, their contribution to the broader educational objectives remains limited.

2.2.1 Additive Factors Model and Performance Factor Analysis

The Additive Factors Model (AFM) [13] is a logistic regression that extends item response theory by incorporating a growth or learning term. The model gives the probability p_{ij}, in log-odds, that a student i will get a problem step j, with related KCs (k) specified by q_{jk}, correct based on the student's baseline ability (θ_i), the baseline difficulty of the related KCs on the problem step (β_k), and the learning rate of the KCs (γ_k). The learning rate represents the improvement on a KC with each additional practice opportunity, so it is multiplied by the number of practice opportunities (T_{ik}) that the student already had on the KC:

$$log(\frac{p_{ij}}{1-p_{ij}}) = \theta_i + \sum_k (q_{jk}\beta_k + q_{jk}\gamma_k T_{ik})$$

Eq 1. Additive Factors Model (AFM).

The Performance Factor Analysis (PFA) [55] is an extension of the AFM model that splits the number of practice opportunities (T_{ik}) into the number of successful opportunities (s_{ik}), where students successfully complete the problem steps, and the number of failed opportunities (f_{ik}), where students make errors. Both (s_{ik}) and (f_{ik}) have their own slopes, γ_k and ρ_k:

$$log(\frac{p_{ij}}{1-p_{ij}}) = \theta_i + \sum_k (q_{jk}\beta_k + q_{jk}\gamma_k s_{ik} + q_{jk}\rho_k f_{ik})$$

Eq 2. Performance Factor Analysis (PFA).

While PFA tends to produce better predictions than AFM, its parameters are not particularly meaningful [43], particularly because their slope interpretation is ambiguous. One interpretation, which is consistent with the intention of PFA, is that these parameters capture individual differences in student mastering that are particular to KCs (i.e. student-KC interactions). Namely, students who make more errors on a KC than otherwise expected will master that KC more slowly than otherwise expected. An alternative, and perhaps more straightforward, interpretation is that the success slope (S-slope; γk) and failure slope (F-slope, ρk) represent different learning rates for prior initially successful versus failed practice opportunities. An indication supporting this notion is the occasional occurrence of a negative F-slope, which, under the second interpretation, can be interpreted as students being unable to learn from unsuccessful attempts [43]. This interpretation could be problematic since it implies that a true novice does not learn (or even unlearns) from making errors. This seems unlikely given modeling and empirical evidence that making errors can contribute significantly to positive learning, as long as feedback is provided [47, 60, 86].

Chapter 3
Meaningful Models

As we discussed in the previous chapters, the recent trend in EDM primarily focuses on developing models that can perform better than the state-of-the-art models on some datasets using various metrics, such as cross-validation errors or AUC. While these works may advance academic benchmarks or stimulate novel developments in machine learning, their real-world significance and practical utility remain questionable without clear evidence of actionable impact. As researchers in the EDM community, our work should aim beyond incremental gains in metrics; ultimately, it should enrich educational practice, support informed decision-making in classrooms, and empower educators to better understand and enhance student learning outcomes. The value of machine learning models lies not merely in their predictive power, but fundamentally comes from its use and how it contributes to stakeholders' benefits. This "use" can manifest as direct deployment in a real-world application, such as adaptive learning platforms that tailor educational content in real-time, or as indirect utilization through the insights it offers to humans, by revealing meaningful patterns and explanations that empower educators, administrators, or policymakers. For example, Klenanov et al. found that students who skip tasks tend to reinforce this behavior over time and create a self-perpetuating cycle, providing an actionable insight: teachers need just-in-time, personalized support to break this cycle [8].

Without bridging the gap from performance metrics to these practical utility, machine learning models risk becoming abstract achievements disconnected from meaningful educational progress. However, simply achieving incremental improvements on benchmark datasets does not automatically translate into meaningful practical outcomes. A model’s predictions alone may solve a task, but the model becomes far more valuable when we actively interpret it—to build user trust or to extract novel insights and inform policy or improve science.

3.1 The Promise of Interpretable Models

How do we transform strong performance scores into real-world value? Researchers in emerging areas such as Human-Centered AI (HCAI) and Explainable AI (XAI) argue that **interpretable** and **explainable** models could be the critical bridge from metrics to practice [26, 32, 36, 48]. Existing studies show that models are far more likely to be adopted when

end-users, domain experts, and regulators can understand, challenge, and act on the outputs. For example, Caruana et al. investigated the evaluation of the application of machine learning for predicting pneumonia risk [11, 16]. The authors report that highly accurate—but opaque—neural-net models "were considered too risky for use on real patients," so clinicians adopted an interpretable generalized-additive model instead Similarly, Darvish et al. employs an exploratory qualitative approach, conducting eleven in-depth interviews to identify key factors influencing XAI adoption in business [19]. They found perceived explainability as a top driver of corporate AI roll-outs, alongside technical readiness and regulatory pressure. These observations raise a key question: what constitutes interpretability in these models?

Interpretability in machine learning is a multifaceted concept, encompassing various perspectives on how and why models should be understandable to humans. At its core, interpretability pertains to the degree to which a human can comprehend the internal mechanics or decision-making processes of a machine learning model. This understanding is crucial, especially in high-stakes domains like healthcare, finance, criminal justice, and education, where decisions can have significant consequences. By elucidating the reasoning behind model outputs, interpretability fosters trust, facilitates compliance with regulatory standards, and aids in identifying and mitigating biases within the models. Moreover, it enables practitioners to diagnose errors, refine model performance, and ensure that the models align with domain knowledge and ethical considerations.

To provide a structured approach to evaluating interpretability, Murdoch et al. propose the Predictive, Descriptive, Relevant (PDR) framework [58], which evaluates interpretability through three key dimensions: predictive accuracy (how well the model predicts outcomes), descriptive accuracy (how well the explanations reflect the model's operations), and relevance (the usefulness of the explanations to a human audience) . This framework underscores that interpretability isn't solely about model simplicity or the availability of explanations but also about the quality and applicability of those explanations to end-users. Additionally, Lipton et al. highlights the ambiguity surrounding the term, distinguishing between transparency—where a model's operations are inherently understandable—and post-hoc explanations, which attempt to elucidate the behavior of complex models after training [26]. Lipton cautions that post-hoc methods may offer insights but can sometimes provide misleading representations of a model's true reasoning.

Further emphasizing the importance of model transparency, Rudin et al. advocates for the adoption of inherently interpretable models, particularly in high-stakes domains where decision transparency is paramount [61]. The authors argue that relying on black-box models, even when supplemented with post-hoc explanations, can be misleading and potentially harmful. They emphasize the importance of designing models that are transparent by design, such as decision trees and scoring systems, which allow stakeholders to directly understand and scrutinize the decision-making process without

the need for additional explanation tools. As an example, Rudin et al. highlight that ProPublica's claim of racial bias in COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) recidivism model [95], a proprietary model that is used widely in the U.S. Justice system for parole and bail decisions, relied on a linear surrogate that poorly represented the recidivism model's nonlinear logic, showing how post-hoc explanations can mislead rather than illuminate [72, 94].

Another important aspect of interpretability is causality. Causal inference techniques are a critical approach to enhance the scientific and practical value of machine learning models. Causal inference explicitly aims to identify cause-and-effect relationships, thereby providing deeper insights into how changes in variables impact outcomes. These causal explanations have the potential to improve decision-making in high-stakes domains by clarifying the consequences of potential actions and interventions. The most rigorous form of causal inference is randomized controlled trials (RCTs), but there are also approaches using naturalistic data without experiments [27] [https://link.springer.com/content/pdf/10.1007/BF00413966.pdf]. Despite its conceptual strength, these causal inference techniques without experiments alone are not sufficient. Accurately establishing causality typically requires original models to use variables and parameters based on related constructs, so the identified relationships are sound causality. Also, closing the loop experiments, often through RCTs, are necessary to validate the identified causality because causal claims derived from observational data can remain speculative, prone to biases, and potentially misleading. Thus, while causal inference enriches the interpretability, it is not sufficient to bridge the gap between machine learning models and real-world values.

3.1.1 Why Post-Hoc Explanations Are Insufficient

Post-hoc methods, such as LIME [22] and SHAP [42], offer flexibility and can be applied to a wide range of complex models, making them appealing for generalization. They are often easier to implement than designing inherently interpretable models that match the predictive performance of black-box models. Despite their utility, post-hoc explanation methods have notable limitations. These methods typically generate feature attributions or surrogate models that highlight which inputs influenced a prediction. They often provide local approximations that may not faithfully represent the model's decision-making process, leading to potential misinterpretations. These explanations can be unstable, with small changes in input data resulting in significantly different explanations, undermining their reliability. Can the post-hoc explanation method provide global approximations? Rudin et al. argues that post-hoc surrogate explanations can never be globally faithful, because any simplification inevitably misrepresents the black-box model in parts of the feature space where their behaviours diverge. If a surrogate were perfectly faithful—matching every input–output pairing—it would replicate the black-box's entire

decision function, so the “explanation” would simply be the original model and inherit its full complexity [71].

Moreover, they often fall short in providing meaningful interpretations that lead to actionable insights. While these methods can highlight which inputs influenced a prediction, they rarely offer guidance on how to modify inputs to achieve desired outcomes. For instance, if a model predicts that a student is at risk of underperforming, a post-hoc explanation might indicate that low prior quiz scores contributed to this prediction. Yet, it doesn’t suggest specific interventions, such as targeted practice on particular concepts or adjusting study strategies, leaving educators without clear, actionable steps. Given these critical limitations, inherently interpretable models emerge as the more suitable solution, particularly in contexts requiring actionable and understandable knowledge. Such models, unlike post-hoc methods, can inherently support simulatability, human-understandable representation, and alignment, so they not only facilitate reliable and robust insights but also provide clear pathways for effective interventions and informed decision-making.

3.2 Interpretability Is Not Enough: The Case for Meaningful Interpretation

Despite widespread recognition of interpretability as a critical factor in machine learning adoption, interpretability alone does not fully address the practical concerns faced by stakeholders. Merely designing models to be inherently interpretable does not, by itself, ensure that stakeholders will engage in interpretation and trust the models, or that their interpretations will translate into genuinely valuable and actionable insights. Interpretability is foundational, yet it must be paired with meaningful interpretations that resonate with stakeholders’ needs.

A direct benefit of a model is its deployment in real-world applications; in high-stakes domains, successful deployment hinges on strong stakeholder trust. However, trust cannot be established simply by labeling a model as interpretable—without carefully examining its actual interpretations, claims of interpretability remain useless. Previous studies show that PFA usually predicts student outcomes accurately, and its parameters have clear and practical meanings, most notably the “success” and “failure” slopes, which represent learning rates after correct answers and errors. Closer inspection, however, reveals a problem: the model often produces negative failure slopes, implying that students “unlearn” after errors [89]. This finding contradicts prior works in learning science, which asserts that making mistakes typically supports, rather than harms, learning [9, 47, 82]. Such interpretations, despite the model's interpretability, can undermine stakeholder trust rather than build it, highlighting that claims of interpretability alone are insufficient. True interpretability means verifying that explanations match expert knowledge and give stakeholders insights they can trust and use.

Another key benefit is the generation of useful insights for stakeholders. A model is valuable only when its findings are both verifiable and useful. In an educational context, teachers and educators rely on predictive models not only to forecast outcomes, but to guide interventions. A prediction without a clear rationale or suggested course of action offers limited practical benefit to those making decisions on the ground. Researchers likewise seek models that uncover new scientific patterns. For example, Koedinger et al. discover that students do not show substantial differences in their rate of learning from fitting statistical growth models to 27 datasets, which challenges the prior hypothesis that different learners acquire competence at different rates [33]. To deliver such insights, models must expose parameters that align with domain-related constructs, knowledge-component difficulty, prior knowledge, and student motivation [18], so researchers and practitioners can meaningfully interpret what the model reveals about the learning process.

Moreover, I argue that simply having interpretable parameters is not enough; these meaningful insights can only be derived from the actual interpretation that emerges from those parameters. For example, Liang et al. applied DiCE counterfactuals to flag features linked to student risk and fed these into LLMs to craft “personalized” feedback. The output, however, amounted to raw data, e.g., “aim for about 332 minutes”, without contextual guidance. This case shows that even when the models are interpretable with understandable parameters, the model's interpretability has no practical value unless they are interpreted and yield actionable insights for educators and learners [37]. This illustrates that even with an interpretable model, failure to engage with the interpretations can result in missed opportunities for intervention and improvement.

This concern is just as relevant for models developed primarily for prediction. For example, even when a model achieves high accuracy in forecasting student outcomes, its practical value is limited without clear reasoning behind those predictions. For instance, if a model identifies a student as likely to fail a course, the prediction alone does not equip educators to intervene effectively. Unless the educator understands what factors are contributing to that risk and why, the prediction fails to support meaningful action. Concrete and actionable interpretations are what enable educators to tailor their interventions, whether by assigning targeted remedial exercises, offering counseling, or addressing non-academic barriers to learning. This is because interpretation is the "why" behind the prediction—the bridge between model output and pedagogical decision-making. Interpretability must therefore extend beyond surface-level explanations to provide context-sensitive insights that align with educators’ needs.

A study by Cohausz et al. on student dropout prediction offers a compelling example of why interpretation matters more than explainability alone [14]. The authors used LIME to identify key features influencing dropout risk, such as poor attendance. While these outputs technically explain the model's decisions, the study found that such surface-level insights were not sufficient to support meaningful intervention. For example, knowing that a student has low attendance may signal risk, but without understanding why

that attendance pattern exists—whether due to disengagement, external obligations, or institutional barriers—educators cannot respond appropriately. The authors argue that true interpretability requires combining these data-driven outputs with domain-specific reasoning, such as theories from education and social sciences. Only then can stakeholders move from knowing what influenced a prediction to understanding how and why it matters, enabling targeted, context-aware actions.

3.2.1 What Interpretation Is and Is Not

It is also important to clarify what interpretation is—and what it is not. Interpretation involves making sense of model outputs in ways that connect to theoretical or practical understanding within a domain. It includes analyzing parameters, identifying underlying patterns, and reasoning through their implications. By contrast, visualization techniques such as heatmaps or feature importance plots are not interpretations in themselves—they are tools that assist in the process of interpretation. Simply displaying which features are most important does not explain why they matter or how they relate to outcomes in context. Without this deeper layer of analysis, even models that appear interpretable risk being reduced to superficial insights.

3.3 Reframing Interpretability: The Meaningful Model

Building on the preceding discussion, it becomes clear that what we ultimately need are models that go beyond surface-level interpretability and support meaningful interpretation in practice. In this section, I propose a framework to conceptualize the notion of "**meaningful models**". In general, a meaningful model consists of two core components: (1) an **interpretable model**, and (2) **meaningful interpretations** that lead to **practical values**, reinforced by **three key properties**: (a) simulatability (b) human-understandable representations (c) alignment with human reasoning and domain theory.

What makes an interpretation meaningful? **An interpretation is considered meaningful when it leads to practical values.** Practical values can be defined in various ways depending on the context and specific objectives. Considered the context of EDM, I propose that practical values can be organized into three types:

> **Trust-building** (for real-world deployment) is critical because, in educational settings or other high-stake domains, stakeholders must trust the model before it can be effectively deployed in real-world applications. For example, adaptive learning systems rely on knowledge tracing models to dynamically tailor educational content to students' real-time performance. However, highly predictive models like DKT are not widely implemented due to their lack of interpretability and

the absence of meaningful parameter interpretations. Conversely, simpler and less predictive models, such as BKT, enjoy broad adoption because their parameters are intuitive and align clearly with human reasoning, thereby fostering trust among educational stakeholders.

Actionable insights are meaningful findings, usually derived from the interpretation of model parameters, that lead to practical recommendations or understandings that stakeholders, such as teachers or administrators, can use to enhance decision-making processes. For example, Liu et al. demonstrate that meaningful interpretations of AFM parameters, such as learning rates for knowledge components' slopes, can lead to new actionable insights [41]. These insights include improved discovery of cognitive models, resulting in practical applications like the redesign of intelligent tutoring systems.

Scientific discovery refers to novel insights that advance our understanding of educational phenomena. Meaningful interpretations can reveal previously unnoticed relationships or patterns, leading to new theories or confirming existing ones within learning science or related fields. For example, Rachatasumrit et al. demonstrate how a simulated learner can serve as a hypothesis-testing engine. They integrated simulation results with empirical evidence to clarify apparent contradictions between the testing effect and worked-example principles. This approach exemplifies how meaningful interpretations from computational models can significantly advance scientific understanding in education.

Having clarified the practical values that meaningful interpretations can offer, the next step is to establish the key properties that constitute a type of model that can lead to those meaningful interpretations.

3.3.1 Simulatability (a)

Simulatability refers to the extent to which a human can mentally trace and comprehend the entire decision-making process of a model. As Lipton et al. highlights, even models traditionally considered interpretable, such as linear regressions, can become opaque if they involve large numbers of parameters or overly complex interactions [40]. This observation challenges the common assumption that simplicity alone guarantees interpretability, emphasizing instead that models must remain cognitively manageable for human users. Sparsity also plays a significant role in enhancing simulatability. Sparse models, characterized by a limited number of non-zero parameters, are generally easier for humans to interpret. For instance, logistic knowledge tracing models, despite potentially having a large number of parameters, frequently employ one-hot encoding for

their input features, resulting in a sparse representation that ensures their simulatability. However, it's important to note that sparsity or simplicity does not always equate to interpretability. Rachatasumrit et al. discuss the simpler PFA model, which, despite having fewer parameters, can be less interpretable compared to its extended version (PFA-h) that introduces additional parameters. The interpretability issues in the PFA models are due to confounded parameters making their practical meanings ambiguous. In contrast, the additional parameters in PFA-h attempted to de-confound those parameters, thereby enhancing clarity and interpretability despite increased complexity [59]. Simulatability can be crucial for meaningful models because only when humans can mentally follow a model's reasoning can they validate its conclusions, trust its outputs, and translate those insights into informed actions.

3.3.2 Human-understandable Representation (b)

Human-understandable representation emphasizes the importance of models presenting their internal workings—such as features, parameters, and structures—in ways that are naturally interpretable to humans. Central to this idea is the selection and use of model features that align closely with human intuition and domain-specific concepts, which significantly facilitates the interpretation of the model's behavior. For example, in educational data mining, features like correctness (whether a student answered a question correctly), and time spent on a task are immediately meaningful to educators. In contrast, abstract features, such as complex embeddings from deep neural networks, can be powerful predictors yet are difficult for humans to directly interpret, as their meanings and implications are unclear. Moreover, parameters within a model should ideally reflect concepts or processes that domain experts can readily recognize and reason about. For instance, logistic knowledge tracing models frequently employ parameters representing the difficulty of knowledge components and the student's prior knowledge, making it straightforward for educators to interpret these parameters and adjust instructional strategies accordingly. Ultimately, prioritizing human-understandable representation ensures models not only make accurate predictions but also provide actionable insights that practitioners can effectively utilize to inform decisions, strategies, and interventions.

3.3.3 Alignment (with Human Reasoning and Domain Theory) (c)

Alignment (with human reasoning and domain theory) ensures that a model's output and their interpretations align closely with established human knowledge and logical reasoning in a specific domain. In practice, this means models should not contradict foundational theories or accepted guidelines within their fields—or, at the very least, any deviations from established knowledge must be clearly explained and justified through these

underlying theories. For example, Zambrano et al. use both recurring self-reports (SR) and classroom observation (BROMP) [53] to measure student emotion during a study and develop two sets of affect detectors corresponding to SR and BROMP-based measures of student emotion. Their analysis shows that SR and BROMP-based detectors are picking up on different signals. Prior studies on the differences between self-report and observational measurements points to the availability of different signals, which supports the findings. For instance, an observer might see a student who has reached an impasse as experiencing confusion or frustration, but self-reported confusion requires some metacognitive recognition on the part of the student.

Unlike simulatability and human-understandable representations, which are properties of the model itself, alignment is a property of its interpretation. Human-understandable representations and alignment are closely related but play distinct roles. In short, human-understandable representations make each component of the model readable, whereas alignment checks that the pattern that those components express coheres with human reasoning or accepted domain theory. Put differently, the former asks, "What does the model reveal?" while the latter asks, "Why does it matter, and what do those representations imply?" A model can expose clear parameters yet still mislead if their learned patterns contradict theory; likewise, aligning with theory is unlikely if the underlying representations are opaque. Truly meaningful models must therefore possess both qualities in tandem.

In summary, a model becomes meaningful not simply because it is easy to understand, but because the understanding it affords leads to something useful—whether that is a trust building, actionable insights, or a deeper scientific discovery. In this sense, meaningful models combine structural interpretability with interpretive engagement. Without the act of interpretation, even the most interpretable model fails to deliver value unless its outputs are examined, interpreted, and connected to the domain context in which it operates.

3.3.4 What about Performance metric

Although this thesis critiques an exclusive focus on performance metrics such as BIC, AUC, or cross-validated accuracy, these measures are by no means useless. From a human-centred standpoint, however, these metrics are not goals in themselves, as they offer stakeholders no direct value. They are *means*, not *ends*. Robust fit statistics or accuracy scores assure us that a model is at least reasonably faithful to the data, and we certainly do not want an interpretable model whose predictions deviate wildly from reality. In this sense, metrics act as a complementary lens: they help us audit and refine models so that any interpretations we draw rest on a sound empirical base.

This perspective differs from the traditional benchmark culture, where the main goal is to outperform the current leader, even by a fraction of a percent. Instead, I treat a good fit as a basic requirement, not a model selection approach. When a model marginally improves accuracy, we must ask: does the gain make a practical difference, or is it functionally negligible? Sometimes marginal improvements do matter. For example, Weitekamp et al. show in his analysis that a small boost in knowledge-tracing accuracy can markedly reduce unnecessary problem assignments in adaptive tutors. However, this is not always true, so model selection must balance predictive performance with the concrete value delivered to stakeholders, adopting a holistic view that keeps metrics in a supportive role.

3.4 Literature Reviews of Meaningful Models

This emphasis on the importance of meaningful interpretations raises a pressing question: to what extent are current models in *educational data mining* and *artificial intelligence in education* research achieving this standard? In the following section, I review recent literature in both fields to highlight a recurring and critical issue—a widespread lack of attention to interpretation, and in many cases, to interpretability itself. Through this review, I aim to show how this gap limits the practical utility and trustworthiness of many models that are otherwise methodologically sound.

In conducting this review, I examined two distinct groups of literature: first, recent studies explicitly claiming interpretability, and second, a selection of papers published in prominent EDM during 2024. Among the works asserting interpretability, a significant majority primarily delivered human-understandable representations—such as meaningful input features or structured outputs—but stopped short of offering genuine interpretations of model behavior or outcomes. Some of these studies employed neural network architectures, inherently limiting simulatability due to their complex structures. In contrast, others leverage simpler, linear models, which naturally enhance simulatability by allowing stakeholders to mentally reconstruct the model's reasoning. Despite these methodological distinctions, a critical gap persisted uniformly across both categories, such that none of the reviewed papers undertook the actual interpretive task—that is, they did not explicitly articulate why certain model decisions or predictions occurred, leaving the essential step from model transparency to meaningful interpretation notably absent.

In reviewing the EDM and AIED proceedings from 2024, I analyzed a total of 75 full papers, not including any studies on which I am listed as co-authors, of which 22 involved the development or training of models. Among these, several explicitly highlighted interpretability as a key contribution; however, upon closer examination, their claims largely amounted to providing human-understandable representations—such as intuitive feature sets, visualizations, or clearly labeled model parameters—without progressing further to substantive interpretation. While these representations indeed

facilitated some degree of human insight, they did not fulfill the deeper interpretative task of explaining why the model made specific predictions or how these decisions could be meaningfully connected to educational theories or practices. In the end, there are 3 papers (4.00%) that actively engage in meaningful interpretations.

Condor et al. specifically addressed interpretability in automatic short answer grading systems by employing Neural Additive Models (NAMs) [15], which integrate the predictive strength of neural networks with the transparency offered by additive modeling. Their approach relied on engineered features grounded in the Knowledge Integration (KI) framework [39], thus providing human-understandable inputs that educators could conceptually grasp. However, given that the model's core still leverages neural network architectures, it inherently lacks simulatability. Although the authors argued that NAMs offer "easily interpretable visualizations of the model's prediction functions," supposedly providing educators insights into student comprehension, these visualizations fell short of generating genuinely actionable insights for teachers. Consequently, the practical value of these visualizations remains limited, highlighting a crucial distinction between superficial transparency and meaningful interpretation.

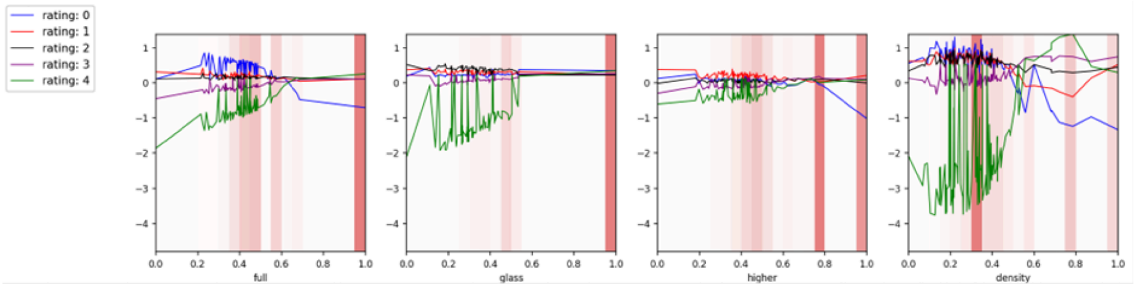

Figure 1. Condor et al.'s visualization of model's prediction function

Another example that clearly emphasizes the importance of meaningful interpretations is Liang et al.'s work on generating personalized feedback based on DiCE counterfactual explanations [37, 49]. The authors employed DiCE to automatically identify parameters indicative of students at risk and subsequently used these parameters as input for large language models (LLMs) to create personalized feedback. Figure 2 provides an example of the feedback generated in their study. However, upon examination, it becomes evident that the feedback primarily consists of a summary of factual data points rather than actionable guidance. For instance, a recommendation such as "aiming for a total of approximately 332 mins" is presented without clear contextual meaning or practical direction. Thus, while the parameters and explanations provided by DiCE are indeed human-understandable, the absence of genuinely actionable interpretations significantly limits their practical utility for educators and students.

	F35(Week7)	F38(Week7)	F32(Week5)	F33(Week5)	Label
Original	0.2	0.1	6	22	**At-Risk**
Counterfactual	332	156	–	–	**No Risk**
Feedback	Dear [Student Name], I hope this message finds you well. I am writing to deliver feedback on your recent performance and offer insights to help you leverage your potential fully. Firstly, I commend your active participation during Week 5, where you accessed the LMS on 6 d and had a total of 22 learning sessions. This level of engagement is commendable as it demonstrates a commitment to your learning journey. Well done on maintaining this consistent connection! However, I've noticed in Week 7 that your time spent on the LMS was only 0.2 min, and your longest learning session lasted just 0.1 min. Engaging deeply with course materials is essential for comprehension and success. To mirror the commitment you showed in Week 5, I suggest aiming for a total of approximately 332 min in the LMS and a continuous learning session of at least 156 min. Allocating time efficiently will give you the opportunity to participate more actively in discussions and solidify your understanding of the material. Could you share what might have caused the decrease in your engagement during Week 7? I believe that by reflecting on this, together, we can identify strategies to assist in managing your study time effectively. Remember, I am here to support you. Please feel free to ask any questions or seek further clarification. Warm regards, [Your Name]				

Figure 2. Liang et al.'s personalized feedback from LLM with DICE.

However, among the reviewed literature, a few studies stood out by genuinely engaging in the task of interpretation. Klebanov et al. developed a model designed to predict instances of students skipping content within an interactive reading application. Their approach leveraged a generalized linear mixed model (GLMM) using clearly interpretable features such as turn duration and turn length. Crucially, their interpretive analysis revealed that prior instances of skipping were the strongest predictor of future skipping behaviors, highlighting the tendency for student disengagement to escalate if not proactively addressed. Furthermore, they illustrated how their model could be practically applied: by estimating an optimal activity duration and enabling personalized adjustments when ongoing disengagement was identified for individual students. This study represents a compelling example of how models built with inherent interpretability can yield direct, actionable insights that meaningfully support educational interventions.

Table 1. Literature review on meaningful interpretations in EDM and AIED.

Paper	Key Properties			Meaningful Interpretations
	Simulatability	Representation	Alignment	
Shi et al. [76]	×	✓	~	~
Hoq et al. [50]	×	✓	×	×
Tsabari et al. [79]	×	✓	×	×
Cao et al. [10]	✓	✓	~	×
Lindsey et al. [38]	✓	×	~	~
Liang et al. [37]	~	✓	~	×
Yu et al. [88]	×	✓	~	×
Klebanov et al. [8]	✓	✓	✓	✓
Condor et al. [15]	~	~	×	~
Rodrigues et al. [67]	~	✓	×	×
Atil et al. [5]	×	×	×	×
Sonkar et al. [77]	×	×	×	×
Tsutsumi et al. [80]	×	✓	✓	×
Queiroga et al. [58]	✓	✓	~	×
Islam et al. [46]	×	~	×	~
Ghanem et al. [24]	×	×	~	×
Demirtas et al. [21]	✓	✓	✓	✓
Alam et al. [52]	×	~	×	×
Zhao et al. [93]	✓	✓	~	~
Zambrano et al. [91]	✓	✓	✓	✓
Acosta et al. [2]	×	~	×	~
Kim et al. [90]	✓	~	~	~

Chapter 4
Good Fit Bad Policy: Why Fit Statistics Are A Biased Measure Of Knowledge Tracer Quality[1]

A popular application of knowledge tracing models is estimating students' mastery of individual KCs to adaptively select subsequent problems based on students' current abilities. Mastery of a KC is typically characterized as the point when a student's predicted chance of correctly answering future question items associated with the KC exceeds some preset mastery threshold, typically chosen in the range 85–95% [4]. Thus, the challenge of knowledge tracing is to actively adapt to students as they practice to optimize their use of time—giving them enough practice problems for each KC to ensure full domain mastery, but not more than this to avoid wasting time better spent practicing new material. Thus, the ideal knowledge tracer jointly minimizes over-practice, the number of prescribed practice problems given after the student has reached mastery, and under-practice, the number of practice problems which a student would still need to solve in order to achieve mastery.

Unfortunately, over- and under-practice are not directly measurable quantities. Instead, the relative quality of knowledge tracers is typically compared on the basis of the overall fit of their underlying student performance models to student data. Overall fit statistics take the form $\pi(\widehat{y}, y)$ and measure the degree to which the continuous student model predictions y are a good approximation of the discrete sequence of binary correctness values $y = y_0, \ldots, y_n$ (correct=1, incorrect=0) collected from student transaction logs. Prior work has used a variety of fit statistics for knowledge tracer comparisons including Mean-Square Error (MSE), prediction accuracy, log-likelihood, AIC [3], BIC [96], and Area under the receiver operating characteristic curve (AUC). In this work, we demonstrate that overall fit statistics can in fact be a biased basis for knowledge tracer comparison since there are circumstances where a model's total predictive performance can be improved without any corresponding change in the behavior of a knowledge tracer utilizing that model. A model can fit better without producing any corresponding reduction in the number of over- and under-practice problems experienced by students.

[1] This work is adapted from Rachatasumrit et al. [62] published at AIED 2024.

This issue directly reinforces the central thesis on meaningful models, arguing that a truly valuable model must prioritize generating actionable insights that guide practical decisions, rather than solely achieving high predictive accuracy. We show through simulation that it is possible for a knowledge tracer model to fit better than a baseline model but perform worse when deployed in practical applications, such as adaptive learning systems which require mastery-based item selection.

4.1 Over-Practice and Under-Practice

Although counts of over- and under-practice are not directly measurable from student data, they can be defined relative to a notion of a student's ground-truth learning curve—their true probability of answering next question items correctly at each practice opportunity. Framed in non-stochastic terms, a student's ground truth curve for a given KC represents the degree to which that KC has been mastered at each learning opportunity. It captures the progression of complex cognitive factors beyond the scope of what statistical performance models typically capture. A point along the curve captures the degree to which a student has partially constructed knowledge—a notion that statistical models typically were estimated solely from binary observations of correct and incorrect performance.

By reference to a ground-truth learning curve and a choice of mastery threshold, a model's instances of under-practice are those where the performance model predicts performance to be above the mastery threshold when the ground truth is below it, and the model's instances of over-practice are those where it predicts performance to be below mastery when the ground-truth is above the mastery threshold (Fig. 3).

Student performance modeling can be framed as estimating students' groundtruth learning curves from the noisy sampling of performance data collected from tutoring system transactions. The logic of comparing knowledge tracers by their overall goodness-of-fit to data is motivated by the idea that an optimal recreation of the ground-truth learning curve should produce an optimal prediction of student mastery. However, this perspective conflates the logic of offline statistical modeling, in which goodness-of-fit can be used to justify hypotheses about students' learning trajectories and their relationship to learning materials, with the narrower aims of online item selection. In this context, a knowledge tracer's purpose is simply to make one critical decision: after a student completes each problem it decides whether to continue prescribing new practice problems with particular KC requirements or not. Thus, certain variations in the predictions of a student performance model simply have no bearing on the real-world quality of their knowledge tracer.

Figure 3 demonstrates how this can be the case by offering an illustration of a hypothetical set of performance model predictions relative to a ground-truth learning curve. The intersection of the ground-truth curve with the mastery threshold divides the

figure into 4 quadrants. Predictions in the top-left and bottom-right quadrants are instances where the model would cause under- or over-practice. The dots and x's in Fig. 1 represent the predictions $\widehat{y}_A$ of a baseline model A. Consider that there is also a comparison model B with predictions $\widehat{y}_B = \widehat{y}_A + \delta$ perturbed by some δ which brings B closer to the ground truth than A. With these perturbations B's expected overall fit to a sample of the ground-truth curve should be better than model A's. However, only a subset of the shown perturbations would produce improvements in mastery prediction, only those perturbations which move predictions out of the over- and under-practice quadrants (e.g. like δ_4 and δ_5).

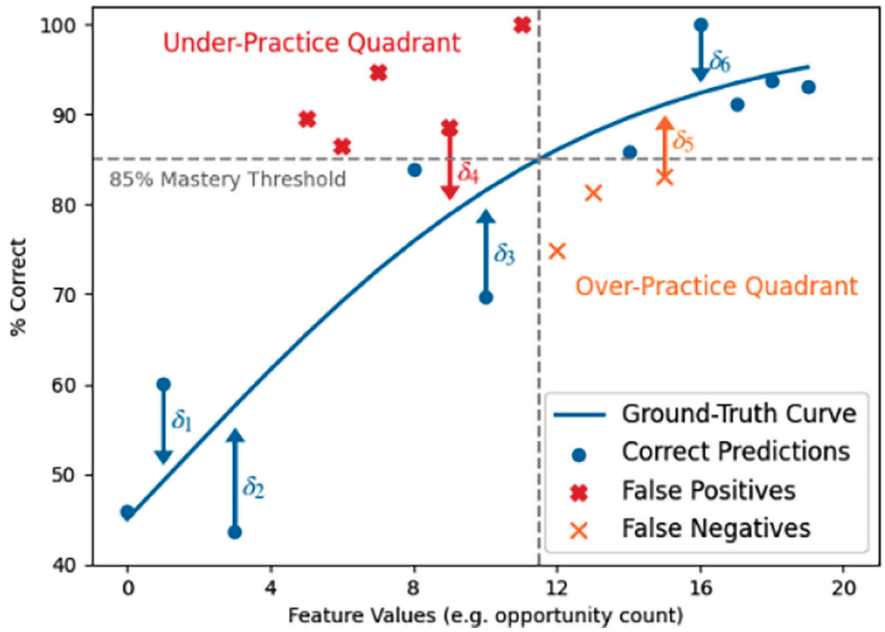

Figure 3. Illustration of over-practice and under-practice attempts

A core hypothesis of this work is that the prediction differences between different types of student performance models mostly do not correspond to differences in expected over- and under-practice like perturbations δ_4 and δ_5. Instead, we hypothesize that the majority of model improvements are like $\delta_1, \delta_2, \delta_3$, and δ_6: inconsequential to levels of over- and under-practice, and generally outside the neighborhood of the ground-truth mastery threshold. One reason to expect this result is that the more data that models have about students the more similar their predictions are likely to be. We expect models to have the

greatest difference in their predictions under uncertain circumstances, particularly in early practice attempts when evidence about the student's knowledge is sparse.

To test this hypothesis we utilize synthetic student data to establish ground-truth learning curves. Then we fit various student performance models on the synthetic data and utilize the ground-truth curves to measure over- and under-practice. We evaluate whether the student performance models which produce the least over- and under-practice are also the best fitting models with respect to overall performance statistics like AUC and MSE. Finally, we graph MSE as a function of ground-truth probability to evaluate whether differences in model fit tend to be greatest within or outside the neighborhood of the mastery threshold.

4.2 Methods

We utilize 3 models for synthetic data generation and evaluation: BestLR [23], DKT [57], and PFA [55]. For each dataset, we use each model to create a simulated dataset and evaluate each generated dataset with all 3 models to create a 3 × 3 experiment. In all cases, we use implementations from Gervet et. al. [23].

Our synthetic data generation works by (1) fitting a generation model to the real data, (2) predicting an error rate for each transaction with a fitted model, and using the predicted value as a ground truth for an error rate in synthetic data, (3) sampling a synthetic outcome for each transaction in the synthetic data based on the corresponding error rate. In this work, we use the same 7 real-world datasets from Gervet et. al. [23], so we generated 21 synthetic datasets for our experiment using 3 generation models. For each synthetic dataset, we use random cross-validation splitting by students. The data of 90% of the students are used for training and the data of the other 10% are reserved for the test set. We resample and retrain 5 times for each condition, examining the relative counts of over- and under-practice on the test set between models, and compare this to their relative AUC scores on the test set. We report the average and standard deviation for each metric across replicates.

4.3 Results and Discussion

Table 2 shows the average instances of over- and under-practice and Table 3 shows the average AUC for each dataset and evaluation model pair. Conventional evaluations assume that between two models the one with the higher predictive performance (e.g. higher AUC) will be the better model—the one expected to make fewer over- and under-practice errors. However, our results demonstrate that this assumption is not always true. We find that in 43% of the synthetic datasets, there are pairs of models where the higher AUC model commits more over- and under-practice errors than the lower AUC

model. These results support the hypothesis that overall fit statistics are not a reliable measure of a knowledge tracer's ability to optimally select next items for students, and challenge the credibility of conventional approaches to comparing knowledge tracers.

4.4 Limitation and Conclusion

One of primary limitations in this work is the ground-truth identification. As in prior works that have utilized synthetic data for analyses of student performance models [61], our method relies upon a theoretical commitment to an underlying model for generating ground-truth curves. Thus our method is not a stand-in replacement for traditional metrics of model fit which evaluate models directly on datasets. Yet, methods which draw comparisons between statistical models and synthetic ground truths have the potential to enable deeper evaluations than the simple notion of what fits best is best.

In this work, we have utilized synthetic data generated by popular knowledge tracers to test whether models with the highest overall fit statistics necessarily produce the best predictions of student mastery. Our method allows us to answer questions of the nature: what is the quality of knowledge tracer X's item selection assuming student learning behaves like model Y? Varying models X, Y, and datasets we find that in 43% of the synthetic datasets, models with higher measures of overall predictive performance (i.e. AUC) were worse than a comparison model with a lower predictive performance at minimizing over-practice and under-practice. We conclude that traditional measures of overall performance (e.g. AUC) are in fact not reliable proxies for rates of over- and under-practice. These results support the core proposition of meaningful models arguing that the value of models are from their use, either to discover actionable insights or be deployed in practical applications. Therefore, metrics on predictive performance are not only insufficient but can also be misleading when it comes to model comparisons.

Table 2. Average numbers of over- and under-practice for each dataset and model

Dataset	Generate	BestLR	DKT	PFA
algebra05	BestLR	4.577 ± 0.235	7.261 ± 0.199	10.843 ± 0.433
	DKT	13.164 ± 5.184	8.300 ± 0.327	32.522 ± 1.957
	PFA	9.067 ± 0.672	13.116 ± 0.835	5.028 ± 0.563
assistments09	BestLR	3.355 ± 0.078	4.488 ± 0.181	5.393 ± 0.151
	DKT	7.280 ± 0.151	4.000 ± 0.107	9.597 ± 0.265
	PFA	4.258 ± 0.136	5.706 ± 0.246	3.309 ± 0.184
assistments15	BestLR	2.398 ± 0.045	4.388 ± 0.323	2.961 ± 0.056
	DKT	8.096 ± 0.107	3.963 ± 0.063	8.233 ± 0.167
	PFA	2.377 ± 0.118	4.997 ± 0.186	2.425 ± 0.043
assistments17	BestLR	2.638 ± 0.045	3.567 ± 0.060	5.297 ± 0.085
	DKT	6.334 ± 0.226	2.808 ± 0.027	3.614 ± 0.098
	PFA	4.663 ± 0.280	4.738 ± 0.395	3.495 ± 0.581
bridge_algebra	BestLR	3.936 ± 0.094	5.494 ± 0.217	6.405 ± 0.132
	DKT	14.033 ± 0.368	6.751 ± 0.165	22.319 ± 0.712
	PFA	4.762 ± 0.300	6.539 ± 0.200	3.759 ± 0.218
spanish	BestLR	2.447 ± 0.022	4.213 ± 0.16	3.173 ± 0.083
	DKT	10.798 ± 0.222	4.600 ± 0.194	12.701 ± 0.345
	PFA	2.397 ± 0.041	4.324 ± 0.145	2.109 ± 0.036
statics	BestLR	3.962 ± 0.205	4.263 ± 0.185	10.559 ± 0.442
	DKT	10.379 ± 0.415	5.095 ± 0.235	19.067 ± 0.843
	PFA	8.333 ± 0.687	7.565 ± 0.589	3.743 ± 0.457

Table 3. Average and SD of AUC for each dataset and evaluation model

Dataset	Generate	BestLR	DKT	PFA
algebra05	BestLR	0.794 ± 0.002	0.728 ± 0.004	0.716 ± 0.004
	DKT	0.808 ± 0.003	0.764 ± 0.007	0.737 ± 0.002
	PFA	0.689 ± 0.004	0.645 ± 0.004	0.705 ± 0.002
assistments09	BestLR	0.712 ± 0.003	0.636 ± 0.006	0.653 ± 0.003
	DKT	0.736 ± 0.005	0.696 ± 0.004	0.670 ± 0.004
	PFA	0.629 ± 0.003	0.565 ± 0.007	0.653 ± 0.003
assistments15	BestLR	0.721 ± 0.005	0.702 ± 0.006	0.713 ± 0.005
	DKT	0.658 ± 0.001	0.674 ± 0.002	0.656 ± 0.001
	PFA	0.659 ± 0.002	0.630 ± 0.001	0.659 ± 0.003
assistments17	BestLR	0.734 ± 0.004	0.717 ± 0.005	0.654 ± 0.004
	DKT	0.702 ± 0.002	0.728 ± 0.001	0.617 ± 0.001
	PFA	0.636 ± 0.002	0.619 ± 0.002	0.639 ± 0.002
bridge_algebra	BestLR	0.834 ± 0.031	0.780 ± 0.033	0.780 ± 0.034
	DKT	0.774 ± 0.003	0.747 ± 0.008	0.705 ± 0.004
	PFA	0.699 ± 0.005	0.645 ± 0.002	0.715 ± 0.003
spanish	BestLR	0.820 ± 0.003	0.764 ± 0.001	0.811 ± 0.004
	DKT	0.808 ± 0.006	0.813 ± 0.006	0.788 ± 0.003
	PFA	0.813 ± 0.006	0.763 ± 0.006	0.814 ± 0.006
statics	BestLR	0.799 ± 0.007	0.785 ± 0.010	0.661 ± 0.010
	DKT	0.804 ± 0.005	0.801 ± 0.004	0.665 ± 0.005
	PFA	0.661 ± 0.005	0.647 ± 0.004	0.670 ± 0.004

Chapter 5
Building Meaningful Models

5.1 Toward Improving Student Model Estimates through Assistance Scores in Principle and in Practice[2]

This work is an example of how we identify an issue with the configuration of an existing model (binary outcomes in AFM), which causes it to not be interpretable in some scenarios and develop a new meaningful model (PC-AFM) that addresses the identified issue. The motivation for this work is that, although common methods such as AFM and BKT perform adequately, they rely on simple right-or-wrong responses. Therefore, they are restricted by using only binary student performance (e.g. correct/incorrect response), which could suffer from an information loss due to its dichotomized nature. In ITS, student performance outcome is often reduced to a binary indicator—first-attempt correct or incorrect—where any error or hint is coded as failure. This binary roll-up simplifies modelling but discards nuanced behaviour captured by ITS. We instead use an Assistance Score, the total number of errors and hints a student produces on each step. Preliminary analysis shows that assistance scores correlate with AFM-predicted error rates, implying they supply additional information beyond simple correctness. In this work, we are interested in whether or not an assistance score model could be a better predictor of a student's change in performance than a dichotomous model like AFM. Particularly, our research questions are: (1) How can we develop an effective statistical measurement model that uses assistance scores? and (2) How do we compare two different response models?

5.1.1 Method

AFM [13] is a logistic regression that extends Item Response Theory by incorporating a growth or learning term. Our extension of AFM to support a polytomous outcome measure, like Assistance Score, is inspired by the Partial Credit Model (PCM) [45], which is an adjacent-categories logit model [81]. The model was designed to work with ordered polytomous response categories with a specific order or ranking of responses, which is

[2] This work is adapted from Rachatasumrit et al. [61] published at EDM 2021.

the case for Assistance Score. It is widely applied in aptitude testing to allow for partial credit for near correctness of a response. In adjacent-categories logit models, we model the odds of a higher category relative to the adjacent lower one, and this paired comparison creates the ordering of the categories.

Assistance Score can be interpreted in the partial credit framework as follows. A student who gets a problem step correct on their first try or after fewer errors or hint requests is more likely to have the associated competence than a student who makes many errors or requests multiple hints before getting the step correct. Thus, students making no errors and needing no hints get full credit (Assistance Score = 0) and students with errors and/or hint requests get partial credit in rough proportion to the number of hints and errors.

The Partial Credit Additive Factors Model (PC-AFM) builds upon these two different statistical models, AFM and PCM. For a student i and a step j, there is a set of probabilities $P_{ij} = \{p_{ij};\ a = 0,\ 1,\ \ldots,\ A\}$ describing the chance for student i to get Assistance Score a on the step j, where A is the maximum Assistance Score. In this work, we decided to limit an Assistance Score to 5 because values above this tend not to be meaningful and rare, but extreme outliers (e.g., where assistance score is over 20 or even 140!) would significantly bias the model. 98% of our data have an Assistance Score of 5 or less. We extend AFM to use a multivariate generalized linear mixed model, and the link function in logistic regression takes the vector-valued form.

$$f_{link}(P_{ij}) = \begin{pmatrix} f_{link,1}(P_{ij}) \\ \ldots \\ f_{link,A}(P_{ij}) \end{pmatrix} = \begin{pmatrix} log(\frac{p_{ij1}}{p_{ij0}}) \\ \ldots \\ log(\frac{p_{ijA}}{p_{ijA-1}}) \end{pmatrix}$$

Eq 3. PC-AFM link function model.

Note that $f_{link,\,0}$ is not included due to the number of nonredundant probabilities. PC-AFM uses adjacent-categories logits as a link function based on PCM. The ath adjacent-categories logit is the logit of getting an Assistance Score a versus a − 1. Each link function is an extended version of AFM's linear model (Eq. 4) with a level parameter (α_a), which represents the difficulty to improve from an Assistance Score a to a−1.

$$f_{link,\,a}(P_{ij}) = \theta\_i + \alpha\_a + \sum_k (q_{jk}\,\beta_k + q_{jk}\,\gamma_k\,T_{ik})$$

Eq 4. Individual PC-AFM link function model.

Inverting this function gives an expression for the probabilities of student i to complete a problem step j with each of the possible Assistance Scores a.

$$p_{ija} = \frac{e^{\lambda_a}}{\sum_{i=0}^{A} e^{\lambda_i}}$$

$$\lambda_a = \begin{cases} 0 \;\; if \; a = 0 \\ \sum_{l=1}^{a} f_{link,\,l}(P_{ij}) \;\; otherwise \end{cases}$$

Eq 5. Probability of a student to complete a step with Assistance Score a.

5.1.2 Experiments

We conduct experiments on both synthetic data and real student data to evaluate the performance of PC-AFM. We used the synthetic data to validate PC-AFM's parameter recovery capability and examine our evaluation strategy in a synthetic environment in which Assistance Score is stochastically derived from student ability alone. In particular, Assistance Scores in the synthetic data are not confounded by other student variations, such as their motivational state. We hypothesized that PC-AFM would work less effectively with the real student data because of non-ability effects on Assistance Score, such as students' help seeking strategies or propensity to game the system.

One of the unique challenges in this work is that goodness-of-fit metrics like BIC are unsuitable for comparing AFM and PC-AFM due to differing outcomes (error rate vs. Assistance Score), despite measuring the same latent construct (student ability). To address this, we employ two methods: evaluating parameter reliability via split-half comparisons, and conducting cross-measure predictions. Split-half comparisons assess parameter consistency, compensating for unknown true parameters in real data. Cross-measure predictions leverage the relationship between binary and polytomous outcomes to identify superior predictive accuracy. As detailed exploration of these methods lies beyond the scope of this thesis, please refer directly to the original work for comprehensive information.

5.1.2.1 Experiment 1: Synthetic Data

In this experiment, we generated six synthetic datasets, varying from 8 to 32 knowledge components (KCs) and 25 to 200 students, to create a controlled environment where

students' performance was derived from Assistance Scores. Using the known student intercepts, KC difficulties, and KC learning rates they sampled an Assistance Score (0 – 5; higher values truncated) for every step, then fit two competing models: the standard AFM trained on binary error rates and the proposed PC-AFM trained on the full Assistance Score distribution. Evaluation focused on three aspects: (1) parameter recovery, the correlation between true and estimated latent parameters; (2) split-half reliability, indicating the stability of estimates across random halves of the data; and (3) predictive accuracy, using under 3-fold cross-validation with random, student-blocked, and item-blocked splits, where each model was required to forecast both outcome types via cross-measure conversion.

Results were consistently in favour of PC-AFM. For parameter recovery, we found that PC-AFM better recovers the true student and KC parameters than AFM in almost all comparisons using correlation (Table 4). The correlations of parameters in split-half comparison are reported in Table 5, which show a similar pattern to the correlation between estimated and true parameters. This demonstrates that the parameter correlation in split-half comparisons, which can be computed in real data, is a reasonable proxy for true parameter recovery, which cannot be computed in real data. Figure 4 illustrates better true parameter recovery using Assistance Score and PC-AFM than using error rate and AFM. PC-AFM parameter estimates (red x's) are generally accurate across the spectrum of known parameter values (x-axis), as can be seen by their closeness to the line, which is identity function (intercept of 0, slope of 1). AFM estimates (blue dots) are generally biased toward the extremes.

Table 4: Correlation between true and estimated parameters in synthetic data.

Dataset	Stu Intercept		KC Intercept		KC Slope	
	PC	AFM	PC	AFM	PC	AFM
KC8_S25	0.978	0.954	0.996	0.802	0.914	0.675
KC8_S50	0.973	0.936	0.998	0.985	0.972	0.964
KC8_S100	0.973	0.931	1.000	0.984	0.952	0.909
KC8_S200	0.975	0.936	1.000	0.979	0.975	0.735
KC16_S50	0.990	0.977	0.998	0.780	0.962	0.933
KC32_S50	0.996	0.988	0.995	0.799	0.929	0.543

Table 5: Correlation between split-halves parameters in synthetic data

Dataset	Stu Intercept		KC Intercept		KC Slope	
	PC	AFM	PC	AFM	PC	AFM
KC8_S25	0.932	0.828	0.990	0.895	0.912	0.498
KC8_S50	0.963	0.906	0.998	0.931	0.972	0.945
KC8_S100	0.980	0.941	0.998	0.850	0.969	0.888
KC8_S200	0.871	0.790	0.999	0.955	0.910	0.894
KC16_S50	0.947	0.857	0.997	0.947	0.927	0.843
KC32_S50	0.967	0.942	1.000	0.883	0.997	-0.345

5.1.2.2 Experiment 2: Real Student Data

In the second experiment, we examine PC-AFM across a variety of real world datasets. We used 6 datasets across different domains (statistics, English articles, algebra, and geometry) from the DataShop repository. For each dataset, we use the KC model that achieves the best BIC reported on the DataShop repository. All KC models coded a single KC per step. The number of KCs ranges from 9 to 64, and the number of students ranges from 52 to 318.

For each dataset, we evaluated both PC-AFM and AFM on 5 independent runs of 3-fold CVs of each type predicting both Assistance Score and error rate. We found that PC-AFM outperforms AFM in Student-blocked in both Assistance Score and error rate CVs in most datasets, which suggests that PC-AFM can achieve better estimates of KC parameters. To validate the hypothesis, we investigated split-halves parameters correlation of both models. We splitted the datasets on students to evaluate KC slopes and intercepts correlation, and we splitted the datasets on KCs to evaluate students' intercepts (Table 6). On average, PC-AFM yields better correlations of both KC intercepts (0.954 vs 0.946) and KC slopes (0.600 vs 0.563), but correlations of student intercepts is significantly higher for AFM (0.784 vs 0.495).

Table 6: Split-halves parameters correlation in real data.

Dataset	Stu Intercept		KC Intercept		KC Slope	
	PC	AFM	PC	AFM	PC	AFM
ds308	0.113	0.486	0.971	0.955	0.745	0.583
ds313	0.490	0.830	0.948	0.937	0.865	0.905
ds372	0.427	0.803	0.985	0.968	0.433	0.639
ds388	0.567	0.873	0.946	0.945	0.225	0.354
ds392	0.830	0.901	0.973	0.964	0.494	0.485
ds394	0.541	0.809	0.904	0.906	0.838	0.413

5.1.3 Discussion

Assistance score should, in principle, improve model parameter estimates and predictions based on them. A student who gets a step correct after just one error or one hint (Assistance Score = 1) is likely to be closer to full acquisition of a KC than a student who makes an error and requests 3 hints (Assistance Score = 4). However, the error rate metric commonly used with BKT and AFM treats these the same, since the student was not correct on their first attempt at the step without a hint. Thus, there is potentially extra information about students' level of knowledge acquisition in the Assistance Score not present in error rate. On the other hand, prior research, for example on gaming the system [6], suggests there are other reasons students may produce repeated incorrect entries or hint requests. These may produce enough confounding variance to make using Assistance Score worse at accurate latent parameter estimation than using error rate.

Assessing whether Assistance Score is a better measure than Error Rate in real student data is complicated in two ways. First, we do not have access to the true parameters in real datasets, so we turn to measures of reliability and predictive validity. Second, we know from models of gaming the system and help seeking that students may produce Assistance Scores for motivational and metacognitive reasons that are potentially independent of a mastery source. In other words, Assistance Scores have a student-driven source of variation that may reduce their effectiveness in estimating student mastery. We hypothesize that our model is struggling to estimate student parameters in the real-world datasets due to variance in students' help seeking behavior.

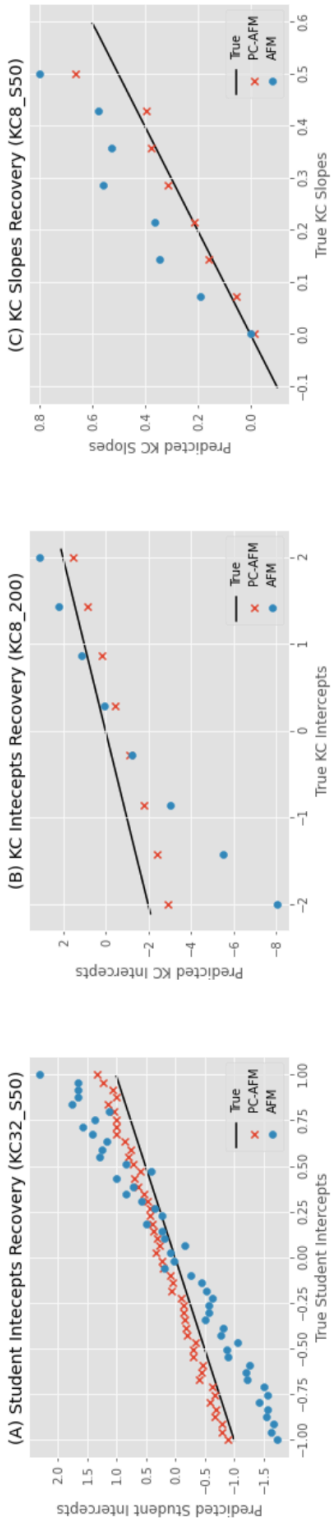

Figure 4. Using Assistance Score and PC-AFM on synthetic data produces better estimates of the true parameters, for all three of student intercepts, KC intercepts, and KC slopes than does using error rate and AFM.

We found that in real world datasets PC-AFM can better estimate KC parameters than AFM, which results in PC-AFM outperforming AFM in Student-blocked CVs. KC parameters estimates significantly impact Student-blocked CVs because they are the sole driver of these predictions. Poor student estimates do not impact Student-blocked CVs because they are not carried from the training to test as blocking means there are different students in the test than training. It does impact Random CVs and Item-blocked CVs because they are likely to have some students showing up in both test and training.

5.1.4 Conclusion

In this work, we investigated whether or not Assistance Score provides a better measurement model than error rate for estimating a student's ability. To pursue this question, we developed a statistical model, PC-AFM, that utilizes Assistance Score. We demonstrated that PC-AFM outperforms AFM when Assistance Scores are synthesized to be meaningful, but its performance is hindered by non-ability variance in students' behavior in the real-world datasets. This work is an example of how we identify an issue with the configuration of an existing model (binary outcomes in AFM), which causes it to not be interpretable in some scenarios and develop a new meaningful model (PC-AFM) that addresses the identified issue. When we applied PC-AFM with real-student datasets, we found that KC parameter estimates are more reliable than student parameter estimates, which led to the insight that the assistance score was heavily influenced by factors beyond their ability, such as motivations. This analysis was only possible because of the interpretable nature of the parameters of PC-AFM, which supports our argument that meaningful models are important in the field of EDM.

5.2 Beyond Accuracy: Embracing Meaningful Parameters in Educational Data Mining[3]

Our goal in this work is to demonstrate that meaningful parameter estimation is not a necessary consequence of more accurate model prediction. We perform this demonstration in the context of two popular models of student learning: PFA [55] and AFM [13]. While PFA tends to produce better predictions than AFM, PFA's parameter estimates are not meaningful because their interpretation is ambiguous. As we will explain in more detail below, interpreting the slope parameters in PFA is difficult because it could mean individual differences in learning rates or differences in prior knowledge or difficulty of

[3] This work is adapted from Rachatasumrit et al. [59] published at EDM 2024.

specific student-KC combinations but it could also mean different learning rates from successful and unsuccessful attempts, or even “unlearning” from errors. Conversely, AFM’s slope is consistently and unambiguously interpretable as learning rate [33]. To demonstrate how PFA’s parameters are confounded, we proposed and evaluated two alternative models (AFMh and PFAh) designed to unconfound the interaction between KCs and students. We demonstrated the capabilities of these alternative models with synthetic data generated from different models and configurations. Then, we conducted an experiment with 27 real-world datasets from Datashop [78], and found that PFA outperforms AFM in 17 datasets, but our further analysis with the new alternative models showed that PFA’s parameters are indeed difficult to interpret. We also argue for the importance of parameter interpretability by comparing AFM and PFA with these alternative models AFMh and PFAh to demonstrate their meaningful interpretations leading to potential insights and applications. In particular, we are interested in these research questions: (1) Can we demonstrate confounding parameters in PFA?, and (2) Do h models have meaningful parameters and also produce better predictions?

5.2.1 AFMh AND PFAh Models

In order to unconfound the student-KC interaction from the success and failure slopes, we need to add additional variables to the models to capture the student-KC interaction. A straightforward approach is to add a variable for each student-KC pair to capture the interaction, but this can lead to overparameterization. Instead, we introduce a success history variable (h_{ik}), which is a ratio between a number of successful past attempts at solving a KC (s_{ik}) and a number of total past attempts at solving that KC (T_{ik}). The intuition behind the success-history variable is that a student who has better prior knowledge of a particular KC would yield higher success rates for the KC. We formulated hik such that its value will be 0.5 at the first opportunity because h_{ik} should be distinguishable in the case of consecutive failed attempts at the beginning. If h_{ik} started at 0, its value would remain 0 regardless of the number of failed attempts at the beginning, which could be problematic for the model:

$$h_{ik} = \frac{s_{ik}+1}{t_{ik}+2}$$

Eq 6. Success-history variable

We incorporated the hik variables into AFM and PFA models to create AFMh and PFAh models, in the term $q_{jk}\eta_k h_k$. The equations for AFMh (Eq. 7) and PFAh (Eq. 8) are below.

$$log(\frac{p_{ij}}{1-p_{ij}}) = \theta_i + \sum_k q_{jk}(\beta_k + \gamma_k T_{ik} + \eta_k h_{ik})$$

Eq 7. Additive Factors Model with History (AFMh)

$$log(\frac{p_{ij}}{1-p_{ij}}) = \theta_i + \sum_k q_{jk}(\beta_k + \gamma_k s_{ik} + \rho_k f_{ik} + \eta_k h_{ik})$$

Eq 8. Performance Factors Analysis with History (PFAh)

5.2.2 Experiment 1: Synthetic Data

In this experiment, we aim to validate the efficacy of our newly developed model in capturing the interaction dynamics between students and KCs. To achieve this, we evaluate this model on synthetic data with known characteristics by sampling model parameters such as student intercepts, KC intercepts, and KC slopes from normal distributions with statistical properties similar to those observed in real-student data. We generated synthetic datasets based on either the AFM or PFA models, serving as the ground truth for student error rates and correctness [61]. To emulate the student-KC interactions observed in real-world scenarios, we introduced variability by augmenting datasets with student-KC interaction effects. This was achieved by sampling values from a normal distribution, reflecting the variance in student performance specific to each KC. Overall, we created 18 dataset groups encompassing varying the number of students (10, 20, and 50), the number of KCs (8, 16, and 32), and the strength of the student-KC interactions (SD = 0.2 and 1.2). We evaluate all four models (AFM, PFA, AFMh, and PFAh) on each dataset. Table 2 and Table 3 show the BIC scores for each model on each dataset in this experiment and summarize the best-fitting models by BIC score.

5.2.2.1 Results

As shown in Table 7, when the student-KC interaction is weak (SD = 0.2), AFM and PFA are the best-fitting models in all datasets depending on the generating model (i.e. AFM is the best-fitting model when the generating model is AFM, and PFA is the best-fitting model when the generating model is PFA). However, when the student-KC interaction is strong (SD = 1.2), the model corresponding to the generation method is the best-fitting model in all datasets, except one (student=10, KC=32, method=PFA+Interaction), as shown in Table 8. In other words, when there is a reasonably strong interaction between students and KCs, the models with the h variable consistently outperform the standard

models. Moreover, the result shows that PFA consistently outperforms AFM when there are student-KC interactions, even when the base generation model is AFM, in which AFMh also consistently outperforms PFA. This supports our hypothesis that PFA parameters are confounded by both the student-KC interactions and two learning rates, but the h variable will be able to unconfound them by capturing the student-KC interactions. Overall, these results also demonstrate the capability of the h models to capture the dynamics of student-KC interactions.

Table 7: BIC scores of all 4 models on synthetic datasets with interaction SD = 0.2. Light grey highlights the best-fitting model among all 4.

Stu	KC	Generation	AFM	PFA	AFMh	PFAh
10	8	AFM	1590.290	1627.946	1598.361	1636.017
		AFM+I	1630.425	1662.996	1634.406	1669.617
		PFA	2091.749	1436.743	1538.479	1444.813
		PFA+I	2072.443	1514.381	1607.171	1522.153
	16	AFM	3818.870	3880.883	3827.027	3885.613
		AFM+I	3808.662	3868.290	3817.426	3877.054
		PFA	4010.223	2807.466	2893.398	2815.151
		PFA+I	3949.252	2840.803	2913.090	2849.557
	32	AFM	6114.022	6196.097	6121.329	6205.297
		AFM+I	6042.236	6125.623	6051.586	6135.080
		PFA	7925.592	6382.965	6676.408	6392.397
		PFA+I	7823.461	6348.209	6673.301	6357.680
20	8	AFM	4791.102	4837.957	4799.797	4846.721
		AFM+I	4601.883	4653.242	4610.647	4662.006
		PFA	6755.818	6403.026	6700.326	6411.790
		PFA+I	6728.999	6445.256	6715.965	6453.907
	16	AFM	6520.145	6597.033	6529.602	6606.491
		AFM+I	6334.954	6405.390	6342.483	6410.950
		PFA	9840.107	8331.947	8969.829	8338.121
		PFA+I	10059.017	8498.802	9050.723	8508.260
	32	AFM	10894.995	10989.292	10905.136	10999.442
		AFM+I	10614.447	10714.491	10624.598	10723.488
		PFA	17967.629	14766.013	15470.549	14776.163
		PFA+I	18373.613	14781.398	15415.666	14791.548
50	8	AFM	7752.478	7813.250	7762.159	7822.930
		AFM+I	7465.130	7529.155	7474.811	7538.835
		PFA	8978.669	6766.349	7572.593	6776.029
		PFA+I	9386.140	7121.818	8032.094	7131.499
	16	AFM	17436.148	17535.014	17446.522	17545.388
		AFM+I	17380.842	17468.669	17390.404	17478.980
		PFA	23980.442	17452.077	19262.037	17462.450
		PFA+I	23881.545	17732.729	19555.968	17743.103
	32	AFM	28246.575	28398.769	28257.642	28409.835
		AFM+I	28505.827	28648.146	28515.574	28658.121
		PFA	33787.825	30985.826	31862.632	30996.893
		PFA+I	35348.852	32002.575	32923.707	32013.642

Table 8: BIC scores of all 4 models on synthetic datasets with interaction SD =1.2. Light grey highlights the best-fitting model among all 4.

Stu	KC	Generation	AFM	PFA	AFMh	PFAh
10	8	AFM	1051.481	1094.670	1059.552	1102.728
		AFM+I	1117.250	1110.974	1095.651	1121.092
		PFA	2086.542	1736.834	1768.927	1744.905
		PFA+I	2442.974	1779.640	1788.976	1778.851
	16	AFM	2209.120	2267.256	2217.864	2276.020
		AFM+I	2412.882	2359.565	2333.930	2359.085
		PFA	3741.063	3585.428	3684.478	3594.192
		PFA+I	4298.942	3809.425	3870.989	3807.412
	32	AFM	6362.627	6444.527	6371.700	6453.985
		AFM+I	7290.315	6785.575	6770.986	6784.784
		PFA	10103.516	8081.974	8434.942	8091.431
		PFA+I	10653.994	8404.126	8559.083	8410.545
20	8	AFM	2387.151	2438.373	2395.171	2447.137
		AFM+I	2811.167	2698.942	2661.740	2695.280
		PFA	5208.531	4508.708	4661.685	4515.641
		PFA+I	5448.687	4676.877	4718.731	4649.611
	16	AFM	5605.182	5687.103	5614.639	5696.560
		AFM+I	6109.225	5905.782	5833.515	5876.967
		PFA	10155.346	7978.861	8504.096	7988.318
		PFA+I	11099.476	8051.809	8196.360	8011.967
	32	AFM	11602.318	11720.229	11612.225	11730.379
		AFM+I	12897.355	11902.381	11796.091	11832.277
		PFA	18625.785	14559.687	15284.133	14569.251
		PFA+I	20953.855	14522.347	14889.161	14501.333
50	8	AFM	9270.245	9337.691	9279.925	9347.372
		AFM+I	10248.059	9472.805	9301.816	9334.143
		PFA	13377.323	10083.043	10708.542	10092.723
		PFA+I	14207.732	9690.340	9895.612	9638.426
	16	AFM	16027.836	16120.648	16038.208	16130.733
		AFM+I	17820.780	16525.557	16326.445	16361.036
		PFA	19711.027	15708.241	16163.369	15718.614
		PFA+I	23266.309	16106.685	16374.813	15996.808
	32	AFM	24554.830	24708.746	24565.897	24719.813
		AFM+I	27686.058	25585.924	25288.177	25326.152
		PFA	47960.208	38961.412	40581.090	38972.479
		PFA+I	52031.370	40238.448	40847.476	40038.740

5.2.3 Experiment 2: Real Student Data

We conducted an experiment with 27 real-world dataset from Datashop across different domains (e.g., geometry, fractions, physics, statistics, English articles, Chinese vocabulary), educational levels (e.g., grades 5 to 12, college, adult learners), and settings (e.g., in class vs. out of class as homework). Table 9. shows the detailed information of each dataset. We evaluated all four models (AFM, PFA, AFMh, and PFAh) on each dataset.

Table 9: Dataset Details, with content domain, grade, number of students, number of observations, and number of KCs.

Dataset	Domain	Grade	EdTech	Student	Obs	KCs
99	Geometry	High	ITS	95	17469	41
115	Chinese	College	ITS	72	19008	248
253	Geometry	High	ITS	41	14785	22
271	Algebra	Middle	Tutor	69	1103	6
308	Statistics	College	Online	52	4152	9
372	English	College	Tutor	99	7128	19
392	Geometry	Middle	ITS	123	41756	38
394	English	College	Tutor	97	5773	13
445	Fractions	Elem	Game	51	4327	21
447	Language	College	Tutor	161	92067	46
562	Fractions	Elem	Tutor	63	48739	102
563	Fractions	Elem	Tutor	64	55407	54
564	Fractions	Elem	Tutor	73	66728	65
565	Fractions	Elem	Tutor	61	57948	78
566	Fractions	Elem	Tutor	58	64025	4
567	Fractions	Elem	Tutor	59	48501	236
1007	CS	College	ITS	49	5063	4
1330	Algebra	Middle	Tutor	2819	39369	24
1387	Fractions	Elem	ITS	84	4032	34
4555	Algebra	High	ITS	129	32125	26

5.2.3.1 Results

Table 10 shows the BIC score of each model on each real student dataset. When comparing between AFM and PFA, PFA outperforms AFM in 17 out of 27 datasets, replicating prior evidence. However, when comparing among all four models, PFA is the best-fitting model in only one dataset (where the difference in BIC score is relatively small), while AFM is the best-fitting model in 4 datasets. AFMh and PFAh are the best-fitting models in 11 datasets each. Among the 17 datasets that PFA outperforms AFM, AFMh is the best-fitting model in 5 datasets. In fact, AFMh outperforms PFA in 24 out of 27 datasets, in contrast to PFAh which outperforms PFA in only 13 out of 27 datasets. Generally, the results demonstrate that the h models usually fit the data better compared to the standard models because they are the best-fitting models in 22 out of 27 datasets.

5.2.4 Discussion

5.2.4.1 RQ 1: Confounding Parameters in PFA

In the synthetic data experiment, we demonstrated the capability of AFMh and PFAh to capture the interactions between students and KCs, as those models outperform standard AFM and PFA when interactions are incorporated in the synthetic datasets. Particularly, PFAh effectively handles the confounding slopes in PFA because the added ηk captures interactions and the slopes capture different rates of learning from errors and successes. It is worth noting that PFA also outperforms AFM in all datasets with strong interactions where the generation method is not AFM without interaction, including AFM with interaction. In other words, PFA is a better fitting model when the generation method includes either student-KC interactions or independent slopes for errors and successes (or both), which attests that the PFA parameters are indeed confounded.

This claim is further validated by the experiment with the real-student datasets. Of the 27 datasets, PFA produces better predictions than AFM on 17 of them – so, indeed, PFA is generally a more predictive model even if it is less interpretable than AFM. However, for 16 of these 17 datasets, either of the new more meaningful models, AFMh (5 out of 17) or PFAh (11 out of 17), yields better predictions than PFA. In other words, PFA is rarely the best-fitting model when we compare it with the models that are designed to separately capture the student-KC interactions. Moreover, even though PFA outperforms AFM in the majority of the datasets, when compared with PFAh and AFMh, it is the best model only in one dataset (6%). On the contrary, AFM is the best model in four datasets (40%). Generally, the results also show that it is possible for a model to be both interpretable and produce better predictions, as evidenced by AFMh and PFAh.

5.2.4.2 RQ2: Meaningful Parameters

We return to the claim that the significance of model parameters and their interpretability supersedes goodness-of-fit or prediction accuracy. The results with real-student datasets demonstrate that AFMh and PFAh are usually better fitting models compared to standard AFM or PFA, but the question remains: do these models hold meaningful interpretations, particularly concerning the h parameter?

It is essential to distinguish between the h_{ik} variable and its associated estimated parameters, η_k. In a meaningful model, parameter estimates typically offer clear interpretations. For instance, in AFM, the student in-tercept represents the student's prior knowledge, while the KC intercept reflects the difficulty of the KC. But what insights does η_k offer?

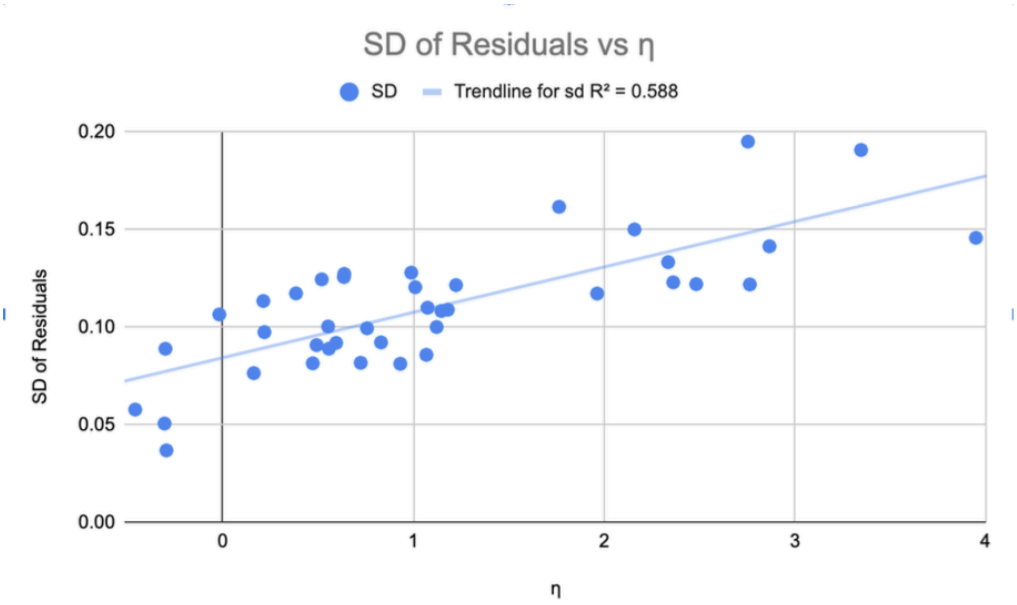

Figure 5. SD of Residuals vs η_k. The residuals and η_k are positively correlated

Table 10. BIC scores of all 4 models on 27 real-student datasets. Light grey highlights a better fitting model between AFM and PFA. Dark grey highlights the best-fitting model among all 4.

DS	AFM	PFA	AFMh	PFAh
99	14568.873	14564.965	14506.087	14522.619
104	6965.241	6978.620	6957.865	6987.335
115	20752.969	20612.962	20722.641	20622.806
253	14598.394	14585.407	14563.883	14585.933
271	1277.940	1305.424	1283.093	1309.691
308	3072.037	3115.442	3079.713	3120.485
1980	6920.579	6944.683	6917.875	6951.888
372	6283.754	6213.442	6207.816	6222.314
1899	5541.982	5555.805	5534.952	5564.308
392	29177.451	29005.429	29006.499	28994.564
394	5580.649	5557.175	5550.959	5565.836
445	4964.794	4971.661	4945.798	4978.275
562	57459.694	56460.229	56410.123	56355.453
563	58377.219	57007.220	56876.034	56840.820
564	67622.473	66165.224	66035.163	65999.477
565	60111.965	57395.729	57057.449	56987.445
566	64040.573	63603.997	63459.030	63470.794
567	49015.532	48010.910	48117.234	48009.947
605	3355.982	3381.284	3361.952	3388.193
1935	8034.666	8052.826	8027.439	8060.300
1330	49749.563	49698.893	49623.904	49622.238
447	87354.605	85040.246	84523.160	84499.571
531	110398.180	106320.620	106032.060	105714.360
1943	127785.500	120277.020	118027.780	117993.150
1387	3298.273	3324.936	3300.726	3330.990
1007	3720.511	3738.319	3688.687	3723.710
4555	36957.404	36506.379	36365.781	36349.639

We investigated the relationship between η_k and the residuals, the difference between the actual outcomes and the model predictions, for each student on corresponding KCs. Particularly, we investigated the ds99 dataset, where η_k ranges from -0.46 to 3.95 (μ = 1.12). Let's first look at the h_{ik} variables. When the KC has a strong variance for the interactions, which means some students are really strong while some students are really weak on the KC, we will also expect a high variance for h_{ik} of that KC. In contrast, when the student-KC interactions have a weak variance, h_{ik} will also be expected to have a low variance. As a result, η_k should be correlated with the variance of the corresponding student-KC interactions. The result from the real-student data, as shown in Fig. 5, supports this hypothesis and shows that the variance of the residuals and η_k are in fact correlated.

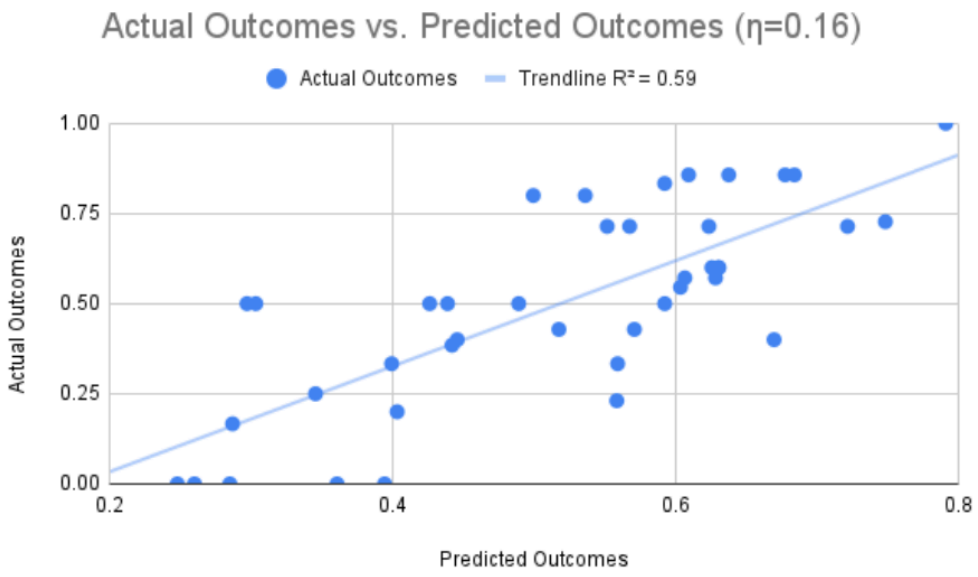

Figure 6. Actual Outcomes vs Predicted Outcomes (η_k=0.16). When η_k is low, students are performing as expected from the model's prediction.

Consequently, the η_k can be interpreted as representing the variance of student-KC interactions of the associated KC. In other words, when η_k is high, some students are really good at the KC while other students are not. For example, number-letter is a KC with a relatively high η_k from the English Article Tutor. The number-letter KC describes a skill that involves selecting an English article (i.e. "a" or "an") to fill in the blank. Examples

of problems with number-letter KC are ”This is the first time that I’ve received ’99’ on a test.” or ”My name begins with ’L’.”. Some, perhaps otherwise struggling, students may learn this skill faster because they happen to focus on the sound of the letter in the following noun and whether it is a vowel or consonant sound. Other, perhaps otherwise good, students may learn this skill slower because they focus on the written letter and whether it is a vowel or consonant. This latter encoding sometimes works, so it is non-trivial to reject in early induction if a learner thinks of it, However, it produces errors and slows down learning overall. On the other hand, when η_k is low, most students are relatively similarly good at that given KC, so the differences in their performance will depend on their overall characteristics, such as student intercepts (prior knowledge). The corollary of this finding is that when η_k is low, students are performing as expected from the model’s prediction (Fig. 6) due to the small variances of residuals. Conversely, students are not performing as expected on the KCs when η_k is large (Fig. 7).

Taken together, these results demonstrate that the h models are not only better fitting models, but their parameters are also meaningful and interpretable. To illustrate the usefulness of the meaningful interpretations, the above suggests a change in the KC model and associated instruction so that the number-letter KC becomes unambiguous and the variance of students’ learning is reduced.

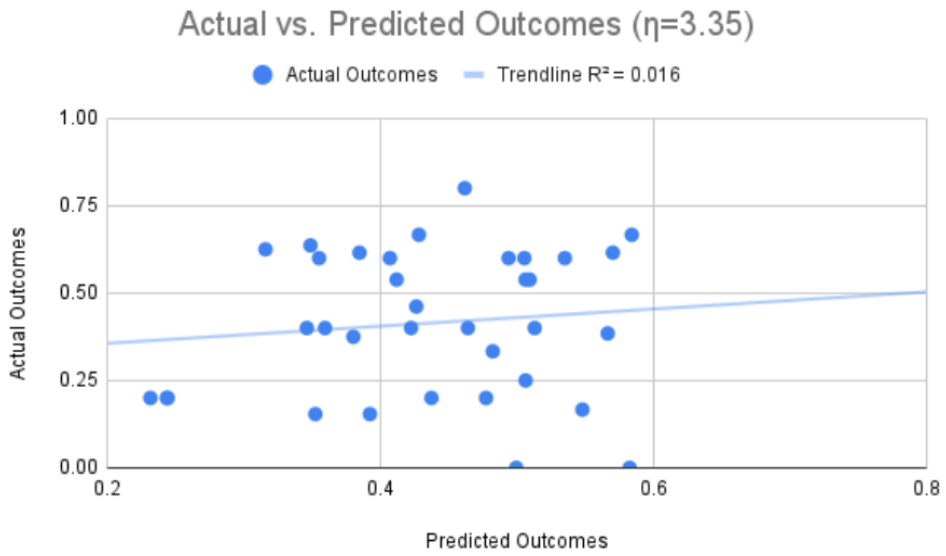

Figure 7. Actual Outcomes vs Predicted Outcomes (η_k=3.35). When η_k is high, students are not performing as expected from the model’s prediction.

5.2.5 Conclusion

In this work, we argued that models with high prediction accuracy do not necessarily exhibit meaningful parameter estimates, which are important for scientific and practical applications. We demonstrated our claim in the context of PFA using both synthetic data and real-student data. The result supported our hypothesis that while PFA is a better fitting model compared to AFM, its parameters' interpretation is ambiguous. Further, we proposed new models AFMh and PFAh, introducing a success-history variable (h_{ik}) designed to capture student-KC interactions, to the existing models. We evaluated their capabilities also with synthetic data and real-student data and demonstrated that the new models are both more interpretable and better fitting compared to PFA.

5.3 Content Matters: A Computational Investigation into the Effectiveness of Retrieval Practice and Worked Examples[4]

This work combines human experiment data with a computational model of human learning to explain why two apparently contradictory instructional techniques: sometimes taking tests (retrieval practice) helps students learn more, while other times studying worked-out examples works better [29, 66, 68, 69]. Previous proposals to address this contradiction, have proposed that problem complexity was the critical dimension that defined whether retrieval practice or worked examples would improve learning [28, 30]. However, as Karpicke et al. pointed out, this explanation does not capture all the evidence [31, 97]. For example, there is ample evidence that retrieval practice improves learning of complex texts [64, 65]]. An alternative hypothesis follows directly from the Knowledge Learning Instruction framework (KLI) [34]. KLI suggests that when learning facts all presented information is critical and should be encoded, whereas when learning skills, only a subset of the presented information is relevant to forming an effective generalized skill. This theoretical proposal is also consistent with procedural differences between research on retrieval practice and example study. Research on retrieval practice generally tests learners' memory of the information presented in repeated trials, whereas research on worked examples generally uses different examples of the same concept in each trial.

To test this hypothesis, Carvalho et al. [12] conducted an experiment using a basic mathematical domain (calculating the area of geometrical shapes) and found a significant interaction between the type of concept studied and the type of training. In this study, we use an AI model of human learning to examine the extent to which the memory mechanism influences human learning and generates such behavioral results similar to Carvalho et al.'s experiment. Specifically, we implemented two models, one without forgetting and another with forgetting. Our results indicate that the simulated learners with

[4] This work is adapted from Rachatasumrit et al. [59] published at AIED 2023 (Best Paper award).

forgetting match human results better and further analysis supports the proposed mechanism. Furthermore, this study provides further evidence for the utility of computational models of human learning in the advancement of learning theory. As proposed by MacLellan et al. [98], the use of such models enables a bridge between learning theory and educational data, allowing for the testing and refinement of fundamental theories of human learning. This study extends this concept by demonstrating the ability of these models to contribute to evaluating theories that can explain even surprising student learning phenomena, for which existing learning theories may offer inconsistent explanations. These computational models are an example of meaningful models that focus on the underlying mechanism, so this work demonstrates how meaningful models can lead to the discovery of scientific insights.

5.3.1 Computational Model of Human Learning

Computational models of human learning aim to precisely characterize how humans learn by constructing artificial (intelligent) systems that interact with simplified learning environments [83, 85, 99]. These models incorporate adaptive learning rules that allow the system to adjust based on interactions with its environment. One notable category of computational models of human learning is represented by simulated learners. Simulated learners are AI systems that learn to perform tasks through an interactive process, such as human demonstrations and feedback, usually with mechanisms that are intended to model how humans learn. In this work, we used the Apprentice Learner framework (AL), a framework for creating simulated learners based on different mechanistic theories of learning. A simulated learner, such as AL, exemplifies meaningful models by offering transparent mechanisms that align with human cognitive processes. This transparency enables these models to produce meaningful interpretations which lead to scientific insights, reinforcing the core thesis of meaningful models. Further details on AL and its operation can be found elsewhere [98, 99]; briefly, AL agents learn a set of production rules through an induction mechanism. The agents receive a set of states as input and search for the existing production rules that are applicable. If none of the existing production rules are applicable, AL agents will request a demonstration of a correct action and go through the induction process to construct a new rule for the current set of states. Later, when the agents encounter states that use the same production rule, the rule will get generalized or fine-tuned according to the examples they encounter. The learning process in AL is largely deterministic but some of the learning mechanisms have stochastic elements. For example, when multiple possible actions are possible, a stochastic probability matching process is used to select which one to execute.

In previous work, AL agents have been shown to demonstrate human-like behaviors in learning academic tasks, such as fractions arithmetic, and multi-column addition [98]. Here, we used AL to test the mechanistic hypothesis that retrieval practice

involves memory and retrieval processes, whereas studying examples involves induction processes. To do this, we developed a memory mechanism in AL and compared the performance of AL agents learning facts and skills in a setup similar to previous empirical results with humans (see also Simulation Studies below). We compare learning outcomes following training of facts and skills, using retrieval practice (practice-only) or worked examples (study-practice). In our study, we employed the same subject matter, but we altered the learning focus between fact acquisition (e.g., “What is the formula to calculate the area of a triangle?”) and skill acquisition (e.g. “What is the area of the triangle below?”).

Additionally, it is crucial to investigate the extent to which memory plays a role in this mechanistic hypothesis. From existing literature, it has been established that retrieval practice has a significant impact on memory. However, the implications of memory processes on the use of worked examples as a learning method are still unclear. For example, what is the potential impact of having perfect memory on these different modes of learning? Therefore, we also created a model without a forgetting mechanism and conducted the same experiment compared to the simulated learners with forgetting.

Our hypothesis predicts that the simulated learners with forgetting will perform similarly to human results, with better performance when learning facts and skills through retrieval practice and worked examples, respectively. The simulated learners without forgetting are not expected to match human results, with retrieval practice being less effective than worked examples in the acquisition of both facts and skills in the absence of a memory mechanism.

5.3.2 Simulation Studies

This work replicates the findings of Carvalho et al.’s [12] experiment on the effect of retrieval practice and worked examples on the different types of knowledge. In their studies, participants were divided into four groups: practice-only training of facts, study-practice training of facts, practice-only training of skills, and study-practice of skills. The participants learned how to calculate the area of four different geometrical shapes (rectangle, triangle, circle, and trapezoid) through a training phase that consisted of studying examples and practicing memorizing formulas or solving problems. To replicate the findings, our pre/post tests and study materials were adapted from the original study.

5.3.2.1 Model Modification

To evaluate our hypothesis that memory and forgetting processes are necessary for a learning benefit of retrieval practice, we leveraged the AL framework to create two models of human learning: a model with forgetting and a model without forgetting (i.e. having a

perfect memory). Our memory mechanism implementation is based on Pavlik et al.'s memory model using ACT-R [54]:

$$m_n(t_{1\ldots m}) = \beta + b_k + ln(\sum_{k=1}^{n} t_k^{-d_k})$$

Eq 9. ACT-R memory model.

In the simulated learners with forgetting, an activation strength (mn) depends on the base activation (β), the strength of a practice type (b_k), ages of trials (τ_k), and decay rates (d_k). The decay rate for each trial depends on the decay scale parameter (c), the intercept of the decay function (α), and the activation strength of prior trials (m_{k-1}):

$$d_k(m_{k-1}) = ce^{m_{k-1}} + \alpha$$

Eq 10. Decay rate model.

The activation strength of each production rule will be updated through the mathematical process described above, every time it is successfully retrieved both through demonstrations/examples or practice testing, but with different corresponding parameter values depending on the type of training. Then, the probability of a successful recall for a production rule will be calculated using the recall equation when simulated learners attempt to retrieve the rule.

5.3.2.2 Study Design

There were a total of 95 AL agents, each agent matching a human participant in [12], assigned to one of four conditions: practice-only training of facts (N = 27), study-practice training of facts (N = 22), practice-only training of skills (N = 18), and study-practice training of skills (N = 28). Each agent went through the same procedure as human participants. It completed 16 pretest questions, 4 study sessions, and then completed 16 posttest questions after a waiting period. In the study session, The agents were divided into two groups: the practice-only group, where they were trained with one demonstration (worked example) followed by three practice tests, and the study-practice condition, where they alternated between both types of training. During the practice tests, the agents were only provided with binary corrective feedback without the correct answer.

5.3.3 Results and Discussion

5.3.3.1 Learning Gain

Similar to the behavioral study in Carvalho et al. [12], we analyzed posttest performance controlling for pretest performance, for each type of trained concept (skills vs. facts) and training type (practice-only, vs. study-practice). A two-way ANOVA was performed to analyze the effects of type of training and type of concept studied on learning gains, and the results showed that there was a statistically significant interaction between the effects of type of training and type of concept in the simulated learners with forgetting ($F(1, 471) = 9.448$, $p = .002$), but none was found in the simulated learners without forgetting ($F(1, 471) = -3.843$, $p = 1$). Moreover, consistent with our prediction, simple main effects analysis showed that the type of training did have a statistically significant effect on learning gains in the simulated learners without memory ($F(1, 471) = 7.364$, $p = 0.007$), but not in the simulated learners with forgetting ($F(1, 471) = 0.845$, $p = 0.359$). On the other hand, the type of concept studied had a statistically significant effect on learning gains in both simulated learners with forgetting ($F(1, 471) = 13.052$, $p < 0.001$) and without forgetting ($F(1, 471) = 29.055$, $p < 0.001$). The similar pattern can also be seen in Fig. 8, comparing the learning gains for each condition between human participants (a), the simulated learners without forgetting (b), and the simulated learners with forgetting (c). The results indicate that the simulated learners with forgetting in a study-practice condition led to higher learning gains for skills than a practice-only condition (19.9% vs 15.8%), $t(228) = -2.404$, $p = 0.009$, but the opposite was true for facts (12.7% vs 15.6%), $t(243) = 2.072$, $p = 0.020$. However, the simulated learners without forgetting led to higher learning gains for both skills (26.1% vs 24.4%), $t(228) = 1.106$, $p = 0.135$ and facts (21.6% vs 20.0%), $t(243) = -1.713$, $p = 0.044$, in the study-practice condition. These results suggest that the simulated learners with forgetting better align with human learning patterns.

5.3.3.2 Error Type

To further investigate the extent to which memory plays a role in this mechanistic hypothesis, we analyzed the types of errors made by simulated learners with forgetting at posttest (since the simulated learners without forgetting cannot commit a memory-based error, it would be unnecessary to conduct the analysis). We classified errors into two categories: memory-based and induction-based. Memory-based errors occurred when an applicable production rule was learned but not retrieved in the final test, whereas induction-based errors occurred when incorrect production rules were found or none were found. Table 11 displays the proportion of each error category, broken down by knowledge type, at the posttest stage. In general, the simulated learners committed more induction-based errors than memory-based errors (56.2% vs 43.8%). Additionally, the

simulated learners in practice-only condition committed less memory-based errors compared to the ones from study-practice condition (41.8% vs 45.7%), but more induction-based errors (58.2% vs 54.2%), $t(466) = -1.467$, $p = 0.072$; even though, both groups exhibited similar proportions of both categories (82.7% vs 83.9% for induction-based errors and 17.3% vs 16.1% for memory-based errors) at pretest.

Table 11. Types of errors (memory-based and induction-based) in posttests for each training condition.

	Memory-based	**Induction-based**
practice-only	41.8% ± 27.9	58.2% ± 27.9
study-practice	45.7% ± 30.3	54.2% ± 30.3
overall	**43.8% ± 29.3**	**56.2% ± 29.3**

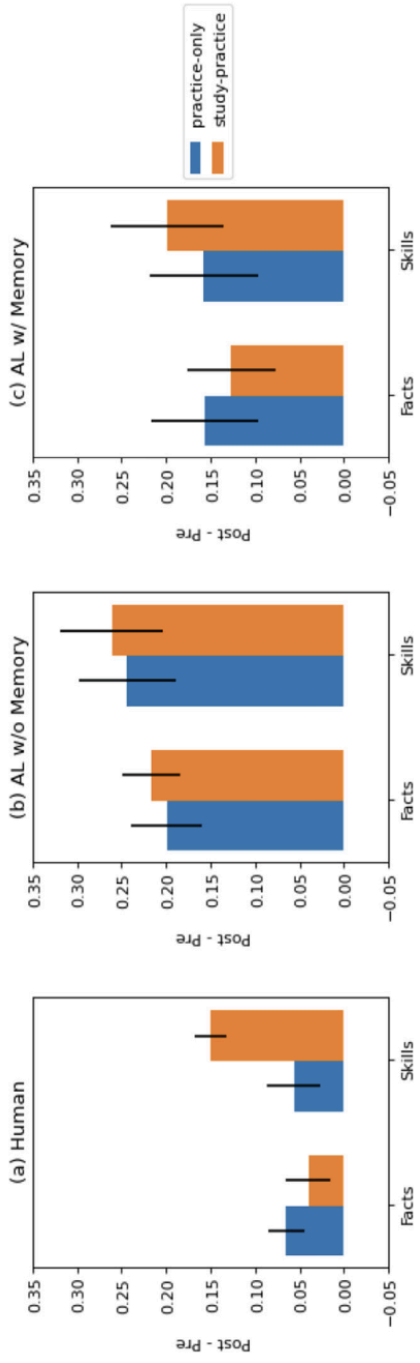

Figure 8. Learning Gains Comparison between type of training and type of concept.

5.3.4 Discussion

Our results indicate that the results of simulated learners with forgetting align well with human results, with retrieval practice being more effective for facts and worked examples being more effective for skills. In contrast, for simulated learners without forgetting, worked examples are more beneficial for both facts and skills, as the lack of a memory mechanism does not allow for the benefits of retrieval practice to be realized. These findings support our hypothesis that, according to the KLI framework [34], retrieval practice improves memory processes and strengthens associations, making it beneficial for learning facts where all presented information is important. Conversely, studying examples improves inference processes and information selection for encoding, making it beneficial for learning skills where only a subset of presented information is relevant.

Interestingly, the introduction of a memory mechanism in simulated learners with forgetting slightly decreases learning gains (22.2% for simulated learners without forgetting vs 13.7% for simulated learners with forgetting), $t(948) = 9.409$, $p < 0.0001$, but does not negate the benefits of worked examples over retrieval practice for skills. Furthermore, the breakdown of error categories revealed more induction-based errors than memory-based errors (59.1% vs 40.9%). This supports our hypothesis that skills learning involves more selectivity and inference, which are better aided by worked examples than by increased memory activation through retrieval practice.

5.3.5 Conclusion

In summary, this study has highlighted the utility of computational models of human learning in bridging the gap between learning theory and data, as demonstrated through examination of unexpected learning phenomena. We started with an unexpected learning phenomena (inconsistencies in the effects of retrieval practices and worked examples on learning), and a proposed plausible mechanism (a mechanism focusing on the selectivity of encoding of the tasks). Then, through the use of computational models, we were able to not only confirm but also examine this proposed learning theory in more depth. Additionally, the ability of these models to examine different learning theories and identify which one best fits human learning, as well as provide valuable insights into the learning process, highlights the potential of computational models in the field of education research. Our findings demonstrate the potential for these models to inform the development of more effective teaching strategies and guide future research in this area.This work also illustrates the benefits of a meaningful model, a simulated learner in this case, that goes beyond simple outcome prediction; by exposing its internal processes, it enables researchers to examine learning mechanisms and draw interpretations that lead to scientific insights.

Chapter 6
Knowledge Tracing Models with Extended Interaction Terms

Traditional statistical thinking has long treated every additional coefficient as a potential liability. Criteria such as BIC [96] impose a penalty on the log-likelihood, so a model with more parameters must improve its fit significantly or be judged inferior. This penalty-based reasoning has shaped an intuition that simplicity and interpretability are two sides of the same coin. This conservative bias, combined with generally limited data availability in educational settings, has made knowledge tracing researchers commonly hesitant to introduce heavy parameterizations, such as student-KC interactions. However, as this thesis argues, the true value of a model lies in its ability to produce meaningful and actionable insights. As we have demonstrated in previous chapters, additional parameters can actually enhance the meaningfulness of a model by yielding interpretations that are theory-aligned and grounded in real-world constructs.

Omitting interaction terms in the models and relying only on main effects may spare a few hundred parameters, but it also eliminates the capability to capture those interactions. These interactions unlock interpretability through personalization: for example, the model may reveal that a given learner consistently excels on a unit overall while stumbling on some specific KCs. Such fine-grained patterns turn the model into a hypothesis-generating engine, guiding educators toward targeted intervention, such as improving instructions to reduce ambiguity. Because each added term is grounded in students' constructs, the extra complexity clarifies rather than obscures, and by capturing meaningful variation it often improves goodness-of-fit as well.

Prior work such as PFA-h incorporated a proxy variable h_{ik}, a ratio of prior successes, that indirectly captures aspects of student-KC interactions. While this approach has been demonstrated to improve upon the original PFA in both goodness-of-fit and interpretability, it remains a coarse approximation. Moreover, preliminary study using synthetic datasets revealed issues with false positives in the PFAh model. Prior work [59] suggests that a given model should be the best performing model (lowest BIC) when the synthetic datasets align with its structure. Specifically, PFAh should perform better in datasets generated with two distinct learning rates for learning opportunities starting with successful or unsuccessful first attempts, whereas AFMh is expected to outperform in

scenarios with a single learning rate. Table 12 presents a comparison of model performance on data simulated using a single learning rate versus data simulated using separate learning rates for learning opportunities starting with successful versus unsuccessful first attempts. While PFAh indeed performed best in scenarios with two distinct learning rates, achieving the lowest BIC in 21 out of 27 cases (77.8%), it unexpectedly also outperformed AFMh in 19 out of 27 datasets (70.3%) under single-rate conditions. These findings demonstrate a limitation in the proxy variable approach utilized by PFAh, suggesting it may not adequately address confounding issues in PFA slopes. To overcome this limitation, we introduced explicit student-KC intercept parameters, sampled from a normal distribution, designed to directly capture the varied patterns of student performance across different knowledge components, thus providing a more accurate reflection of authentic educational scenarios.

Rather than relying on historical aggregates, the student-KC interaction formulation captures how different students respond to different KCs in distinct ways. The explicit student-KC interaction parameters enable the model to surface more interpretable patterns, such as which skills are consistently difficult for subgroups of students, and facilitates richer diagnostic feedback and hypothesis generation. These interaction parameters may also take away the confound producing the false positives in Table 12 so that the success and failure learning rate estimates are more trustworthy. We also directly challenge the assumption that adding a large number of parameters to knowledge tracing models is inherently undesirable. We demonstrate that when these parameters are carefully selected, particularly when they are theory-driven and semantically meaningful, they can enhance interpretability and yield more useful, actionable insights. Importantly, this added complexity does not necessarily degrade model fit; in many cases, it performs comparably to more constrained models while offering greater transparency. Particularly, our research questions are: (1) Do the additional student–KC interaction parameters harm goodness-of-fit? (2) Do they improve the interpretability and practical usefulness of the model?

Table 12: BIC scores of all AFMh and PFAh on synthetic shadow datasets. Light grey highlights the best-fitting model. Green letters indicate the correct best-fitting model, and red letters indicate incorrect ones. PFAh incorrectly fits better than AFMh on simulated data with student-KC interactions and a single slope (left columns).

Dataset	1-Slope		2-Slope	
	AFMh	PFAh	AFMh	PFAh
99	15366.19864	15088.76773	15308.14363	14939.32591
104	6941.984143	6882.42514	6965.888966	6836.709074
115	22047.70432	22042.16587	21822.9575	21673.25315
253	15242.26575	14963.30213	14920.88224	14508.15749
271	1367.028756	1376.242096	1339.102362	1358.980815
308	3163.173181	3184.474983	3186.739676	3191.725551
1980	7236.983122	7136.024606	7380.119404	7207.101123
372	6766.625326	6696.207466	6953.461395	6551.603659
1899	5967.097015	5940.366942	5831.453676	5792.351656
392	33304.18309	32579.90588	33244.59668	31766.27396
394	5942.727325	5898.638637	5903.435371	5774.353689
445	5240.115383	5190.359332	5125.309633	5072.476659
562	57787.44912	56610.90243	56254.36293	53664.00318
563	59340.22199	58027.52844	57541.61146	54371.58408
564	68197.68778	66601.51491	65604.86612	62315.13861
565	62098.36106	60882.94146	58357.49018	54329.66586
566	63558.84719	62193.96991	61979.39471	59431.69133
567	51439.65416	50884.2949	50776.86447	48890.24718
605	4133.607239	4092.715099	3942.951703	3912.214081
1935	8431.453495	8416.575083	8496.135018	8435.192653
1330	49556.72307	49116.32011	49869.50265	49295.53863
447	91613.9094	89467.10344	90561.69605	83628.85426
531	112678.8035	109665.2872	107142.9504	96863.82479
1943	132585.5578	129639.0037	126958.1042	116269.7656
1387	3714.840565	3696.946979	3704.351411	3704.365121
1007	4051.29575	4021.622167	4118.819954	4042.392068
4555	36747.60412	36106.78892	35310.5411	33771.65147

6.1 Extended Models with Interaction Terms

Unlike AFMh and PFAh, which utilize the success-history variable (h_{ik}) as a proxy for student-KC interactions, we propose directly modeling student-KC interactions. Specifically, we introduced unique student-KC intercepts (λ_{ik}) to explicitly capture the interaction between student i and KC k. These intercepts were introduced as random effects to reduce the chance of overfitting and integrated into the AFM and PFA models, creating interaction-based extensions, AFMi and PFAi. The equations for AFMi (Eq. 11) and PFAi (Eq. 12) are provided below. The AFMi model gives the probability p_{ij}, in log-odds, that a student i will get a problem step j, with related KCs (k) specified by q_{jk}, correct based on the student's baseline ability (θ_i), the baseline difficulty of the related KCs on the problem step (β_k), the learning rate of the KCs (γ_k), and the strength of student-KC interaction (λ_{ik}). PFAi is an extension of the AFMi model that splits the number of practice opportunities (T_{ik}) into the number of successful opportunities (s_{ik}), where students successfully complete the problem steps, and the number of failed opportunities (f_{ik}), where students make errors. Both (s_{ik}) and (f_{ik}) have their own slopes, γ_k and ρ_k:

$$log(\frac{p_{ij}}{1-p_{ij}}) = \theta_i + \sum_k (q_{jk}\beta_k + q_{jk}\gamma_k T_{ik} + q_{jk}\lambda_{ik})$$

Eq 11. Additive Factors Model with Interactions (AFMi)

$$log(\frac{p_{ij}}{1-p_{ij}}) = \theta_i + \sum_k (q_{jk}\beta_k + q_{jk}\gamma_k s_{ik} + q_{jk}\rho_k f_{ik} + q_{jk}\lambda_{ik})$$

Eq 12. Performance Factors Analysis with Interactions (PFAi)

We conducted two experiments, on synthetic data and real student dataset, to evaluate the performance and interpretations of newly proposed models (AFMi and PFAi) compared to the standard models (AFM and PFA) and models from prior work (AFMh and PFAh). We used Bayesian information criterion (BIC) as the primary metric to compare model goodness-of-fit. In our experiments, we leveraged 27 real-world datasets from Datashop across different domains (e.g., geometry, fractions, physics, statistics, English articles, Chinese vocabulary), educational levels (e.g., grades 5 to 12, college, adult

learners), and settings (e.g., in class vs. out of class as homework). Table 9. shows the detailed information of each dataset.

6.2 Synthetic Data Experiment

In this experiment, we aim to validate the effectiveness of our newly proposed model in capturing the interaction dynamics between students and KCs and in hopes of eliminating the false positives produced by the PFAh model (see Table 12) so that the success and failure slopes are more scientifically interpretable. We evaluate the model on synthetic datasets designed with known properties. We constructed four synthetic datasets based on two key factors: (1) whether there are different learning rates for successful and failed opportunities, and (2) whether student-KC interactions are present. In the model, these two factors are represented by the number of KC slopes and the inclusion or exclusion of student-KC intercepts, forming a 2x2 experimental design. To reflect the learning rate factors in our synthetic datasets, we used two generating models: AFM, which assumes a uniform learning rate (i.e., a single slope), and PFA, which assumes distinct learning rates for successful and unsuccessful student attempts (i.e., two slopes). For datasets with student-KC interactions, we introduced student-KC intercepts generated by sampling from a normal distribution to capture the varied performance patterns students exhibit across different KCs, mirroring authentic educational scenarios.

Rather than generating data completely from scratch, we strive for our synthetic dataset to accurately reflect the variability inherent in real student data. To achieve this, we created a synthetic “shadow” dataset by resampling student outcomes based on real transaction data. This method aims to ensure that our synthetic dataset authentically represents the variability and complexity observed in real-world student interactions. Our approach to generating synthetic shadow datasets involved three concise steps: (1) collecting students’ and KCs’ parameters from the real student dataset with AFM, (2) predicting error rates for each transaction using the corresponding generating models, and (3) sampling synthetic outcomes for each transaction in the synthetic dataset based on these predicted error rates. In step 2, the generating models depend on the factors discussed above, resulting in four variations for each dataset:

(a) Single KC slope with interaction
(b) Single KC slope without interaction
(c) Two KC slopes with interaction
(d) Two KC slopes without interaction

Since KC parameters were derived exclusively using AFM, a modification was necessary for the two-slope conditions. To clearly distinguish between success and failure slopes and avoid ambiguity relative to the single-slope condition, which frequently occurred when

directly obtaining two slopes using PFA, we applied an offset derived from the average difference between success and failure slopes observed in PFA across all KCs in 27 real student datasets. Specifically, the success and failure slopes were generated by respectively adding and subtracting this offset (≅1.9) to the single KC slope from AFM. This adjustment provided sufficient differentiation between the slopes for robust interpretation. In total, this process resulted in 108 distinct datasets (4 variations x 27 original datasets), and we evaluated all 6 models on each dataset to assess their performance. We hypothesize that in variations with student-KC interactions, the newly proposed interaction models (AFMi and PFAi) will be the best performing models. On the other hand, in variations without student-KC interactions, we expect the standard models to outperform the other models.

6.2.1 Results and Discussion

In contrast to the false-positive issues observed in the PFAh model (see Table 12), Table 13 presents the comparison between AFMi and PFAi on the same problematic shadow datasets, a variation with a single learning rate and student-KC interactions. Whereas earlier analyses revealed unexpectedly better performance by PFAh in single learning rate conditions, the current results suggest that these false-positive occurrences have been substantially reduced. Consistent with our hypothesis, AFMi demonstrated better performance compared to PFAi across all evaluated datasets.

Next, we compare the newly proposed interaction models with the models from our prior work and standard models. Table 14 shows comparison among AFM-based models (AFM, AFMh, and AFMi) on the dataset with a single learning rate. Aligned with our hypothesis, the results show that AFM consistently outperforms the other models in all dataset variations without student-KC interactions. Also, in datasets where interactions were explicitly incorporated, AFMi consistently outperformed both AFM and AFMh. This finding indicates that AFMi might be more effective in capturing student-KC interactions compared to the other models, including AFMh. Similarly, Table X compares the PFA-based models (PFA, PFAh, and PFAi) across dataset variations with two learning rates. Consistent with our hypothesis, PFA performed best in all variations without student-KC interactions. However, in variations with explicitly incorporated student-KC interactions, PFAi demonstrated superior performance, outperforming PFA and PFAh in 15 out of 22 datasets (68.2%), while PFA and PFAh performed best in 4 and 3 datasets respectively.

Furthermore, Table 15 examines the average differences in BIC scores between the best-performing model and the second-best model, grouped by the best-performing model. Raftery et al. suggests that a BIC difference of 10 corresponds to a Bayes factor of approximately 150, which represents very strong evidence of better model fit [63]. The results show that, when interaction models (PFAi and AFMi) achieve the best

performance, the average BIC differences are notably large (228.27 for PFAi and 761.39 for AFMi), whereas the average differences when other models perform best are less than 10

The synthetic shadow datasets experiments showed that extending the model with student–KC interaction terms, despite the substantial increase in parameters, did not diminish goodness-of-fit and frequently outperformed simpler standard models. However, as emphasized throughout this thesis, while fit metrics are valuable for assessing how well a model represents data, the primary benefit for stakeholders often comes from the meaningful interpretations. In the next section, we further evaluate the newly proposed interaction models on real student datasets to determine their effectiveness in producing meaningful interpretations and actionable insights.

Table 13: BIC scores of AFMi and PFAi on synthetic shadow datasets with a single learning rate and interactions. Light grey highlights the best-fitting model.

Dataset	AFMi	PFAi	Dataset	AFMi	PFAi
99	14917.85480	14967.41932	563	57462.86757	57544.34243
104	6806.93481	6841.28230	564	65866.81201	65992.01640
253	14712.44342	14755.18338	565	60275.19152	60362.41424
271	1358.11588	1376.26527	566	61189.52979	61296.21893
308	3111.14526	3158.13974	567	50767.44313	50845.24097
1980	7088.74298	7109.31427	605	4008.47046	4041.44404
372	6665.41712	6695.84614	1935	8279.11886	8314.79350
1899	5926.42428	5940.99247	1330	49320.58286	49163.23575
394	5861.17717	5888.47463	1387	3689.56708	3699.59428
445	5144.25429	5168.92526	1007	3955.77742	3987.69884
562	56287.64314	56384.60297	4555	35568.56610	35600.87201

Table 14: BIC scores of all AFM-based models on synthetic shadow datasets with a single learning rate and interactions. Light grey highlights the best-fitting model.

Dataset	AFM	AFMh	AFMi
99	16003.16932	15366.19864	14917.85480
104	7112.60045	6941.98414	6806.93481
253	16212.81337	15242.26575	14712.44342
271	1379.17088	1367.02876	1358.11588
308	3182.18868	3163.17318	3111.14526
372	6968.71957	6766.62533	6665.41712
392	34656.42379	33304.18309	32153.12604
394	6072.78139	5942.72733	5861.17717
445	5360.97343	5240.11538	5144.25429
447	96662.45519	91613.90940	88519.38245
562	60392.25056	57787.44912	56287.64314
563	63397.69287	59340.22199	57462.86757
564	72608.01544	68197.68778	65866.81201
565	65594.58038	62098.36106	60275.19152
566	68430.13262	63558.84719	61189.52979
567	52988.08207	51439.65416	50767.44313
605	4162.23232	4133.60724	4008.47046
1007	4170.74638	4051.29575	3955.77742
1330	50446.41912	49556.72307	49320.58286
1387	3774.64843	3714.84057	3689.56708
1899	6047.83391	5967.09702	5926.42428
1935	8609.99915	8431.45350	8279.11886
1980	7481.29472	7236.98312	7088.74298
4555	38920.36348	36747.60412	35568.56610

Table 15: BIC scores of all PFA-based models on synthetic shadow datasets with two learning rates and interactions. Light grey highlights the best-fitting model.

Dataset	PFA	PFAh	PFAi
99	15016.74540	14939.32591	14797.27621
104	6863.33347	6836.70907	6823.25745
253	14631.64216	14508.15749	14320.15647
271	1357.55663	1358.98082	1364.00006
308	3188.54034	3191.72555	3164.50216
1980	7212.21023	7207.10112	7184.86711
372	6546.91785	6551.60366	6553.56912
1899	5788.51764	5792.35166	5797.01208
394	5778.14498	5774.35369	5774.47947
445	5074.05619	5072.47666	5054.71567Δ
562	53799.38982	53664.00318	53491.40941
563	54685.72824	54371.58408	53833.03310
564	62569.01138	62315.13861	61827.28138
565	54652.72596	54329.66586	53959.71413
566	59936.52909	59431.69133	58574.47914
567	48953.31052	48890.24718	48890.75109
605	3926.57770	3912.21408	3874.75999
1935	8476.61787	8435.19265	8343.33583
1330	49315.54300	49295.53863	50482.33383
1387	3698.62143	3704.36512	3706.84331
1007	4062.96857	4042.39207	4005.86134
4555	34109.20281	33771.65147	33347.21865

Table 16. Average BIC differences for each winning model type.

Best Model	Δ BIC	Best Model	Δ BIC
PFA	7.267	AFMi	761.394
PFAh	6.878	AFM	8.203
PFAi	228.265	AFMh	N/A

6.3 Real Student Data Experiment

Having established that the models behave as intended in the synthetic experiment, we applied the newly proposed interaction models to real student datasets to further evaluate their capability to produce meaningful interpretations and insights. We applied four models (AFM, PFA, AFMi, and PFAi) across 27 original student datasets without any modifications. Model performance was assessed and compared using the BIC. It should be emphasized that the synthetic data experiment demonstrated how the best performing model is usually determined by the underlying structure of the dataset, particularly the uniformity of learning rates and the presence of student-KC interactions in our case. For example, when AFM is the best performing model, the dataset likely shows little distinction between learning from successes and learning from errors (i.e. success and failure slopes) and only weak student–KC interaction effects. Conversely, datasets in which PFAi emerged as the best model typically implied asymmetry between success and failure slopes and strong student–KC interactions. Given that these factors are not controlled in the real student data, fit metrics like BIC serve as diagnostic tools rather than the primary objective. and the true value of the interaction models lies in the nuanced and informative interpretations they enable.

6.3.1 Result and Discussion

Figure 9 summarizes the number of datasets in which each model achieved the best performance based on a goodness-of-fit metric (BIC). Interaction models generally provided a better fit, with PFAi being the best performing model in 11 datasets and AFMi in 6 datasets. In comparison, the standard AFM model performed best in 7 datasets, while the standard PFA model was optimal in 3 datasets. Table Y provides a detailed breakdown, including specific BIC scores for each model-dataset pair.

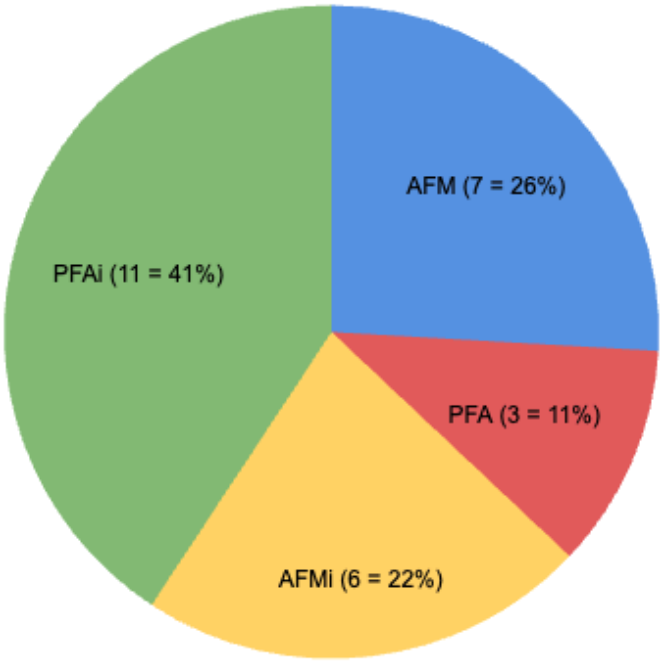

Figure 9. Number of datasets in which each model achieved the best performance.

Experiments with synthetic shadow datasets showed that adding student–KC interaction terms, despite markedly increasing parameter count, did not harm goodness-of-fit and even outperformed the simpler baseline models often. A similar trend appears in the real-student datasets; however, the question remains whether interaction models yield meaningful interpretations.

There is an unresolved debate in learning science on whether learners benefit more from successes or failures during practice. Metcalfe et al. demonstrated that making incorrect guesses followed by clear feedback leads to stronger learning than simply reviewing correct answers [47]. In contrast, research on errorless learning consistently suggests that avoiding mistakes altogether can result in better learning outcomes [9]. Prior studies on this topic typically involve limited numbers of participants and are restricted to single-domain laboratory settings. Since we have demonstrated that student-KC interaction parameters remove a known confound from the PFA's success and failure slopes, we can systematically investigate this question with a large number of datasets. From the result with 27 real student datasets, we found 14 datasets where PFA and PFAi are best fit models, and among these datasets, success slope is greater than failure slope in 12 of them (Table 17). Among them, the average success slope is 0.21 and the average failure slope is 0.04. Across all the datasets, the average success slope is 0.20 and the average failure slope is 0.06. These results provide support for the hypothesis that students learn more from a first-attempt success than a first-attempt failure followed by

feedback. This result demonstrates the benefit of meaningful models for contributing to scientific discovery.

Table 17: Success slope is greater than failure slope in 12 out of 14 datasets where PFA or PFAi are best fit models.

Datasets	Best Model	Success > Failure
115	PFA	✓
372	PFA	✓
392	PFAi	✓
447	PFAi	✓
531	PFAi	✓
562	PFAi	✓
563	PFAi	✓
564	PFAi	×
565	PFAi	✓
566	PFAi	✓
567	PFAi	✓
1330	PFA	×
1943	PFAi	✓
4555	PFAi	✓

Another straightforward interpretation of student-KC intercepts is the degree to which a learner's performance on a particular KC deviates from expectations given their overall performance from student's intercept and KC's parameters. If a KC is overall harder, *all* students will have lower intercept for that KC. Similarly, if a student is particularly well-performing, they will have higher student intercepts, which affect all KCs. For example, a strong student-KC interaction might reveal that a particular student, otherwise well-performing in a geometry course, consistently underperformed on a trapezoid-related KC. This observation can lead to a personalized intervention, such as assigning targeted practice on trapezoid-area problems or offering an alternative explanation for the student.

Table 18: BIC scores of all 4 models on real student datasets with two. Light grey highlights the best-fitting model.

Dataset	AFM	PFA	AFMi	PFAi
99	14568.87265	14564.96489	14516.59703	14542.66767
104	6965.24099	6978.62004	6962.66504	6983.45562
115	20752.96889	20612.96170	20635.73836	20622.63643
253	14598.39431	14585.40750	14561.31853	14583.56381
271	1277.93983	1305.46005	1284.21386	1312.24741
308	3072.03686	3115.44179	3080.04133	3122.33743
372	6283.75412	6213.44189	6233.02525	6219.84040
392	29177.45122	29005.42867	29039.06492	28995.43929
394	5580.64917	5557.17483	5552.41225	5565.11122
445	4964.79420	4971.66119	4949.02413	4978.87248
447	87354.60486	85040.24621	85066.02784	84810.49712
531	110398.18930	106320.62840	105974.78640	105680.95600
562	57459.69442	56460.22911	56728.80081	56413.04766
563	58377.21857	57007.22005	57158.78302	56898.31646
564	67622.47256	66165.22368	66246.13668	65875.75440
565	60111.96550	57395.72894	57326.20781	57115.50833
566	64040.57265	63603.99746	63835.84921	63602.78869
567	49015.53231	48010.90971	48362.67831	48001.16786
605	3355.98173	3381.28360	3361.38202	3389.46702
1007	3720.51097	3738.31900	3697.51079	3727.27133
1330	49749.56307	49698.89304	49760.13279	49709.46276
1387	3298.27292	3324.93566	3305.94790	3333.04327
1899	5541.98172	5555.80526	5545.67621	5564.30834
1935	8034.66627	8052.82604	8043.24955	8061.88567
1943	127785.50560	120277.02760	119222.31420	118483.13860
1980	6920.57900	6944.68258	6922.61692	6953.47619
4555	36957.40390	36506.37913	36459.69693	36396.25938

Beyond analyzing single student–KC pairs, the interaction intercepts can be also aggregated along either dimension to yield higher-level insights. Aggregating across KCs for an individual student can highlight the idiosyncratic pattern of the student. Aggregating across students for a given KC exposes an insight on the KC. For example, a high variance of the intercepts indicates that some learners excel while others lag on that KC regardless of overall ability. In other words, high-achieving students may encounter unexpected difficulty with the KC, whereas lower-performing students can sometimes master it more rapidly. This implies potential ambiguity or prerequisite gaps in the item design. Conversely, a low variance suggests that most students perform as expected from their overall proficiency, implying a well-scaffolded, unambiguous skill. Together, these aggregates translate raw parameters into actionable information for curriculum refinement and targeted support

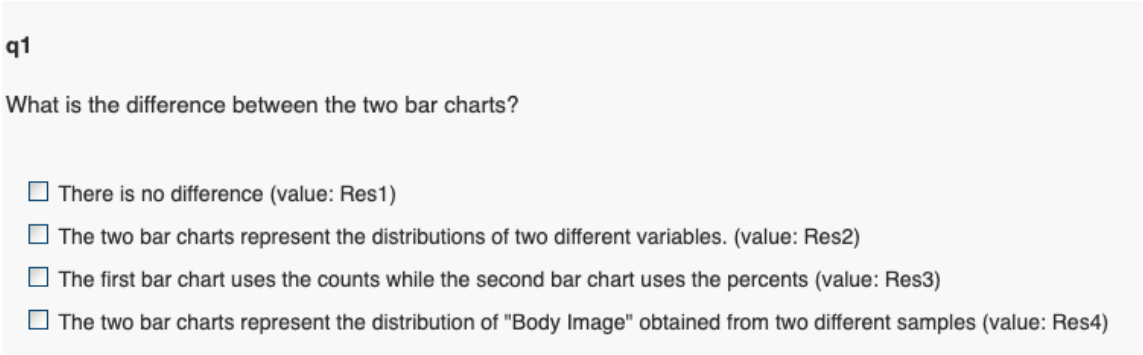

Figure 10. A problem tagged with *skill1*interpbarchart* KC.

Within our real-student dataset, two cases illustrate how the variance of student–KC interaction intercepts signals the clarity—or ambiguity—of a skill. DS115 (Chinese-character domain) contains KCs that map single characters to their meanings or corresponding sound; the task is essentially factual recall, which are less likely to have strong individual variabilities. The maximum standard deviation across student intercepts is just 0.018, indicating that nearly all students perform in line with their overall proficiency and ambiguity is unlikely. In contrast, DS308 (statistics domain) shows an average standard deviation of 0.105. Consider two representative KCs below. The problem with skill1×interpBarChart (SD=0.046) asks learners to identify the apparent difference between two bar charts, whereas a problem with skill1×m1o5 (SD=0.105) is less obvious and requires judging whether a point is an “outlier,” a notion open to interpretation. Moreover, some of the problems with this particular KC are true/false items which encourages guessing and yields wide variability—some high-achieving students might misinterpret the concept, while lower-achieving students may answer correctly by chance.

High variance thus flags KCs whose wording or instructional framing needs revision; reducing ambiguity should narrow the spread of interaction effects and promote more consistent learning.

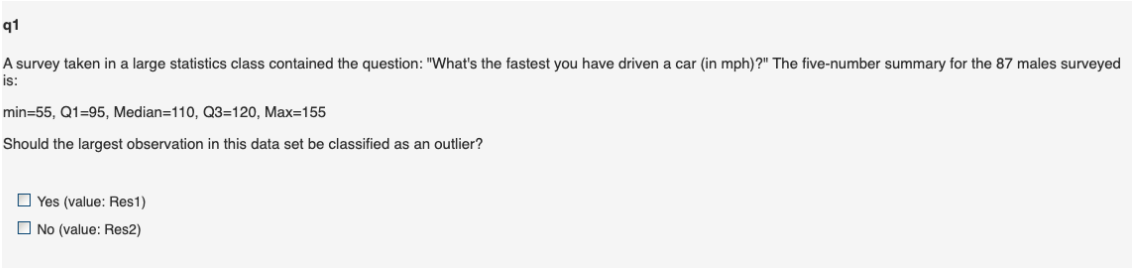

Figure 11. A problem tagged with for *skill1*m1o5* KC.

Prior work has used the PFA-h η parameter (cumulative prior success ratio), which can be interpreted as a proxy for a variance of student–KC interactions. While useful, η is still an approximation. With PFAi, we can calculate the actual variance of student-KC interactions for each KC. Figure 12 plots the absolute η values against the standard deviations of student-KC interaction intercepts derived from PFAi, and the correlation between them is 0.57, which indicates a moderately strong association. Beyond treating η as a rough proxy, the full student-KC interaction intercepts enables us to pinpoint students exhibiting the largest KC-specific deviations and to search for common characteristics that could inform targeted interventions. For example, in the English-article dataset several KCs have significant variance. After discussion with the domain expert who collected the dataset, they hypothesized that this pattern is attributed to first-language transfer: learners whose native language lacks grammatical articles struggle considerably more than those whose language includes them. Unfortunately, the dataset contains only unique student IDs without more information, we cannot directly validate this hypothesis.

Taken together, our findings underline the importance of meaningful models that the real value of a model lies in the insight it provides, not in how few parameters it has or how well it scores on a single fit metric. When the extra terms we add are tied to meaningful learning concepts, such as the way each student interacts with each skill, they maintain solid predictive accuracy while opening a clearer window into student thinking. In this holistic view, the model stops being just a score-predictor and becomes a practical tool that helps teachers spot problems, tailor support, and improve the curriculum.

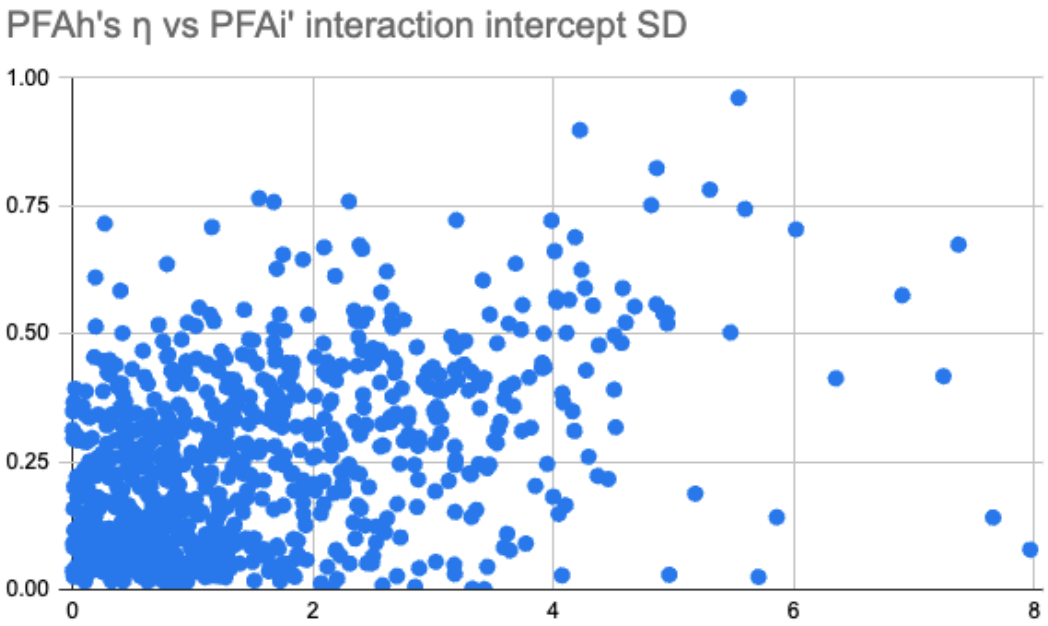

Figure 12. Scatter plots between PFAh's η and PFAi's interaction intercept SD.

6.4 Conclusion

This chapter shows that enriching knowledge-tracing models with carefully chosen student-KC interaction terms produces models that are both empirically sound and pedagogically meaningful. Through experiments on synthetic shadow dataset and real-student datasets, we found that interaction models, like AFMi and PFAi match or exceed the performance of their simpler based models while offering a far richer interpretive capability. Crucially, our findings suggest that parameter count alone is irrelevant when extra terms capture meaningful learning structure, they do not hurt, and often help with interpretations.

These results reinforce the thesis of Meaningful Models: interpretive value, not parameter count or predictive metrics, should guide model choice. When additional parameters are theory-aligned and semantically transparent, they help the model to be more meaningful and lead to actionable insights that are useful for stakeholders.

Chapter 7
Conclusion

In this thesis, I attempt to rethink the model usability, especially in the context of Educational Data Mining (EDM) from a human-centered perspective. It advanced two main claims. First, a model's value lies in what stakeholders can do with it—trust it, act on it, or learn from it—not in marginal gains on fit statistics. Second, making a model interpretable is necessary but insufficient; only when interpretations are actively examined and aligned with domain reasoning do they become meaningful. Building on the literature on interpretable machine learning, I conceptualize the notion of "meaningful models", interpretable models which are actively interpreted for meaningful interpretations that lead to practical values, reinforced by three key properties: simulatability, human-understandable representations, and alignment with human reasoning and domain theory. Researchers need to ensure that their variables and parameters are aligned one-to-one to causal constructs. Researchers also must validate that the results of their models are aligned with domain theory, potentially using synthetic data.

To support these claims, the thesis contributed both conceptual and empirical work. Conceptually, it introduced the notion of a meaningful model and articulated how the three properties jointly enable practical value—trust-building, actionable insights, and scientific discovery. Empirically, it shows through multiple examples of building meaningful models that relying solely on performance metrics can mislead model selection. Across synthetic and real-world datasets, these studies demonstrated that models designed for interpretability can yield relatively good fit metrics or sometimes exceed their less interpretable alternatives while offering more values through their interpretations.

High prediction accuracy has long been the focus of EDM research, but without meaningful interpretation, a model has limited value. To be useful to stakeholders, it is crucial that we prioritize interpreting models to extract insights that can lead to improvements in practical applications or contribute significantly to scientific discoveries. For example, in the learning sciences, it is critical that interpretations result in refining pedagogy or producing insights about learning. Without tangible scientific discovery or practical recommendation, our best models cannot be considered meaningful and are effectively useless. By foregrounding meaningful interpretation alongside stakeholder needs, this thesis offers conceptual guidance and examples for building models that both perform well and make a genuine difference through their interpretations. I hope future

work embraces this broader perspective so that advances in statistical and machine learning models and machine learning models translate into real benefits for stakeholders.

References

1. Abdelrahman G, Wang Q, Nunes B (2023) Knowledge Tracing: A Survey. ACM Comput Surv 55:224:1-224:37. doi: 10.1145/3569576

2. Acosta H, Lee S, Mott B, Bae H, Glazewski K, Hmelo-Silver C, Lester J (2024) Multimodal Learning Analytics for Predicting Student Collaboration Satisfaction in Collaborative Game-Based Learning. pp 224–235

3. Akaike H (2011) Akaike's Information Criterion. In: Lovric M (ed) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg, pp 25–25

4. Arlin M (1984) Time, Equality, and Mastery Learning. Rev Educ Res 54:65–86. doi: 10.2307/1170398

5. Atil B, Sheikhi Karizaki M, J. Passonneau R (2024) VerAs: Verify Then Assess STEM Lab Reports. In: Artificial Intelligence in Education: 25th International Conference, AIED 2024, Recife, Brazil, July 8–12, 2024, Proceedings, Part I. Springer-Verlag, Berlin, Heidelberg, pp 133–148

6. Baker R Why Students Engage in "Gaming the System" Behavior in Interactive Learning Environments

7. Baker RSJD, Corbett AT, Aleven V (2008) More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. In: Woolf BP, Aïmeur E, Nkambou R, Lajoie S (eds) Intelligent Tutoring Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 406–415

8. Beigman Klebanov B, Weeks J, Sinharay S (2024) To Read or Not to Read: Predicting Student Engagement in Interactive Reading. In: Olney AM, Chounta I-A, Liu Z, Santos OC, Bittencourt II (eds) Artificial Intelligence in Education. Springer Nature Switzerland, Cham, pp 209–222

9. Bridger EK, Mecklinger A (2014) Errorful and errorless learning: The impact of cue–target constraint in learning from errors. Mem Cognit 42:898–911. doi: 10.3758/s13421-014-0408-z

10. Cao M, Jr PIP, Chu W, Zhang L (2024) Integrating Attentional Factors and Spacing in Logistic Knowledge Tracing Models to Explore the Impact of Training Sequences on Category Learning

11. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N (2015) Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, pp 1721–1730

12. Carvalho P, Rachatasumrit N, Koedinger KR (2022) Learning depends on

knowledge: The benefits of retrieval practice vary for facts and skills. In: Proceedings of the Annual Meeting of the Cognitive Science Society

13. Cen H, Koedinger K, Junker B (2006) Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement. In: Ikeda M, Ashley KD, Chan T-W (eds) Intelligent Tutoring Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 164–175

14. Cohausz L (2022) Towards Real Interpretability of Student Success Prediction Combining Methods of XAI and Social Science. p 361

15. Condor A, Pardos Z (2024) Explainable Automatic Grading with Neural Additive Models. In: Artificial Intelligence in Education: 25th International Conference, AIED 2024, Recife, Brazil, July 8–12, 2024, Proceedings, Part I. Springer-Verlag, Berlin, Heidelberg, pp 18–31

16. Cooper GF, Aliferis CF, Ambrosino R, Aronis J, Buchanan BG, Caruana R, Fine MJ, Glymour C, Gordon G, Hanusa BH, Janosky JE, Meek C, Mitchell T, Richardson T, Spirtes P (1997) An evaluation of machine-learning methods for predicting pneumonia mortality. Artif Intell Med 9:107–138. doi: 10.1016/s0933-3657(96)00367-3

17. Corbett AT, Anderson JR (1994) Knowledge tracing: Modeling the acquisition of procedural knowledge. User Model User-Adapt Interact 4:253–278. doi: 10.1007/BF01099821

18. Dang SC Exploring Behavioral Measurement Models of Learner Motivation

19. Darvish M, Kret KS, Bick M (2024) An Explorative Study on the Adoption of Explainable Artificial Intelligence (XAI) in Business Organizations. In: Disruptive Innovation in a Digitally Connected Healthy World: 23rd IFIP WG 6.11 Conference on e-Business, e-Services and e-Society, I3E 2024, Heerlen, The Netherlands, September 11–13, 2024, Proceedings. Springer-Verlag, Berlin, Heidelberg, pp 29–40

20. Delibašić B, Vukićević M, Jovanović M, Suknović M (2012) White-box or black-box decision tree algorithms: which to use in education? IEEE Trans Educ 56:287–291

21. Demirtas MA, Fowler M, Cunningham K (2024) Reexamining Learning Curve Analysis in Programming Education: The Value of Many Small Problems. pp 53–67

22. Garreau D, Luxburg U (2020) Explaining the Explainer: A First Theoretical Analysis of LIME. In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. PMLR, pp 1287–1296

23. Gervet T, Koedinger K, Schneider J, Mitchell T (2020) When is deep learning the best approach to knowledge tracing? J Educ Data Min 12:31–54

24. Ghanem B, Fyshe A (2024) DISTO: Textual Distractors for Multiple Choice Reading Comprehension Questions using Negative Sampling. pp 6–17

25. Gillani N, Eynon R, Chiabaut C, Finkel K (2023) Unpacking the “Black Box” of AI in

education. Educ Technol Soc 26:99–111

26. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2019) Explaining Explanations: An Overview of Interpretability of Machine Learning

27. Glymour C, Scheines R (1986) Causal modeling with the TETRAD program. Synthese 68:37–63. doi: 10.1007/BF00413966

28. van Gog T, Kester L (2012) A test of the testing effect: acquiring problem-solving skills from worked examples. Cogn Sci 36:1532–1541. doi: 10.1111/cogs.12002

29. van Gog T, Paas F, van Merriënboer JJG (2006) Effects of process-oriented worked examples on troubleshooting transfer performance. Learn Instr 16:154–164. doi: 10.1016/j.learninstruc.2006.02.003

30. van Gog T, Sweller J (2015) Not New, but Nearly Forgotten: the Testing Effect Decreases or even Disappears as the Complexity of Learning Materials Increases. Educ Psychol Rev 27:247–264. doi: 10.1007/s10648-015-9310-x

31. Karpicke JD, Aue WR (2015) The Testing Effect Is Alive and Well with Complex Materials. Educ Psychol Rev 27:317–326. doi: 10.1007/s10648-015-9309-3

32. Khosravi H, Shum SB, Chen G, Conati C, Tsai Y-S, Kay J, Knight S, Martinez-Maldonado R, Sadiq S, Gašević D (2022) Explainable Artificial Intelligence in education. Comput Educ Artif Intell 3:100074. doi: 10.1016/j.caeai.2022.100074

33. Koedinger KR, Carvalho PF, Liu R, McLaughlin EA (2023) An astonishing regularity in student learning rate. Proc Natl Acad Sci 120:e2221311120. doi: 10.1073/pnas.2221311120

34. Koedinger KR, Corbett AT, Perfetti C (2012) The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. Cogn Sci 36:757–798. doi: 10.1111/j.1551-6709.2012.01245.x

35. Kung C, Yu R (2020) Interpretable Models Do Not Compromise Accuracy or Fairness in Predicting College Success. In: Proceedings of the Seventh ACM Conference on Learning @ Scale. ACM, Virtual Event USA, pp 413–416

36. Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, Sesing A, Baum K (2021) What Do We Want From Explainable Artificial Intelligence (XAI)? -- A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. Artif Intell 296:103473. doi: 10.1016/j.artint.2021.103473

37. Liang Z, Sha L, Tsai Y-S, Gašević D, Chen G (2024) Towards the Automated Generation of Readily Applicable Personalised Feedback in Education. In: Olney AM, Chounta I-A, Liu Z, Santos OC, Bittencourt II (eds) Artificial Intelligence in Education. Springer Nature Switzerland, Cham, pp 75–88

38. Lindsey RV, Shroyer JD, Pashler H, Mozer MC (2014) Improving Students' Long-Term Knowledge Retention Through Personalized Review. Psychol Sci 25:639–647. doi: 10.1177/0956797613504302

39. Linn MC (2000) Designing the Knowledge Integration Environment. Int J Sci Educ 22:781–796. doi: 10.1080/095006900412275

40. Lipton ZC (2018) The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16:31–57. doi: 10.1145/3236386.3241340

41. Liu R, Koedinger KR (2017) Closing the Loop: Automated Data-Driven Cognitive Model Discoveries Lead to Improved Instruction and Learning Gains. J Educ Data Min 9:25–41

42. Lundberg SM, Lee S-I (2017) A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.

43. Maier C, Baker RS, Stalzer S (2021) Challenges to Applying Performance Factor Analysis to Existing Learning Systems. In: Proceedings of the 29th International Conference on Computers in Education

44. Mandalapu V, Gong J, Chen L (2021) Do we need to go Deep? Knowledge Tracing with Big Data

45. Masters GN (1982) A rasch model for partial credit scoring. Psychometrika 47:149–174. doi: 10.1007/BF02296272

46. Md Mirajul Islam, Xi Yang, John Hostetter, Adittya Soukarjya Saha, Min Chi (2024) A Generalized Apprenticeship Learning Framework for Modeling Heterogeneous Student Pedagogical Strategies. doi: 10.5281/ZENODO.12729786

47. Metcalfe J (2017) Learning from Errors. Annu Rev Psychol 68:465–489. doi: 10.1146/annurev-psych-010416-044022

48. Molnar C, Casalicchio G, Bischl B (2020) Interpretable Machine Learning -- A Brief History, State-of-the-Art and Challenges. pp 417–431

49. Mothilal RK, Sharma A, Tan C (2020) Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, pp 607–617

50. Muntasir Hoq, Brusilovsky P, Akram B (2023) Analysis of an Explainable Student Performance Prediction Model in an Introductory Programming Course. doi: 10.5281/ZENODO.8115693

51. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci 116:22071–22080. doi: 10.1073/pnas.1900654116

52. Nazia Alam, Behrooz Mostafavi, Sutapa Dey Tithi, Min Chi, Tiffany Barnes (2024) How Much Training is Needed? Reducing Training Time using Deep Reinforcement Learning in an Intelligent Tutor. doi: 10.5281/ZENODO.12729806

53. Ocumpaugh J, Baker RS, Rodrigo MMT Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual

54. Pavlik PI, Anderson JR (2008) Using a model to compute the optimal schedule of practice. J Exp Psychol Appl 14:101–117. doi: 10.1037/1076-898X.14.2.101

55. Pavlik PI, Cen H, Koedinger KR (2009) Performance Factors Analysis – A New Alternative to Knowledge Tracing. In: Artificial Intelligence in Education. IOS Press, pp 531–538

56. Pavlik PI, Eglington LG Automated Search Improves Logistic Knowledge Tracing, Surpassing Deep Learning in Accuracy and Explainability

57. Piech C, Bassen J, Huang J, Ganguli S, Sahami M, Guibas LJ, Sohl-Dickstein J (2015) Deep Knowledge Tracing. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.

58. Queiroga EM, Santana D, da Silva M, de Aguiar M, dos Santos V, Mello RF, Bittencourt II, Cechinel C (2024) Anticipating Student Abandonment and Failure: Predictive Models in High School Settings. In: Olney AM, Chounta I-A, Liu Z, Santos OC, Bittencourt II (eds) Artificial Intelligence in Education. Springer Nature Switzerland, Cham, pp 351–364

59. Rachatasumrit N, Carvalho P, Koedinger K (2024) Beyond Accuracy: Embracing Meaningful Parameters in Educational Data Mining. pp 203–210

60. Rachatasumrit N, Carvalho PF, Li S, Koedinger KR (2023) Content Matters: A Computational Investigation into the Effectiveness of Retrieval Practice and Worked Examples. In: Wang N, Rebolledo-Mendez G, Matsuda N, Santos OC, Dimitrova V (eds) Artificial Intelligence in Education. Springer Nature Switzerland, Cham, pp 54–65

61. Rachatasumrit N, Koedinger KR (2021) Toward Improving Student Model Estimates through Assistance Scores in Principle and in Practice. International Educational Data Mining Society

62. Rachatasumrit N, Weitekamp D, Koedinger KR (2024) Good Fit Bad Policy: Why Fit Statistics Are a Biased Measure of Knowledge Tracer Quality. In: Olney AM, Chounta I-A, Liu Z, Santos OC, Bittencourt II (eds) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky. Springer Nature Switzerland, Cham, pp 183–191

63. Raftery AE (1995) Bayesian Model Selection in Social Research. Sociol Methodol 25:111–163. doi: 10.2307/271063

64. Rawson KA (2015) The Status of the Testing Effect for Complex Materials: Still a Winner. Educ Psychol Rev 27:327–331. doi: 10.1007/s10648-015-9308-4

65. Rawson KA, Dunlosky J (2011) Optimizing schedules of retrieval practice for durable and efficient learning: how much is enough? J Exp Psychol Gen 140:283–302. doi:

10.1037/a0023956

66. Renkl A (2005) The Worked-Out Examples Principle in Multimedia Learning. In: The Cambridge handbook of multimedia learning. Cambridge University Press, New York, NY, US, pp 229–245

67. Rodrigues L, Avila-Santos AP, Silva TE, da Penha RS, Neto C, Challco G, dos Santos EL, Souza E, Guerino G, Vieira T, Marinho M, Macario V, Bittencourt II, Dermeval D, Isotani S (2024) Knowledge Tracing Unplugged: From Data Collection to Model Deployment. In: Olney AM, Chounta I-A, Liu Z, Santos OC, Bittencourt II (eds) Artificial Intelligence in Education. Springer Nature Switzerland, Cham, pp 91–104

68. Roediger HL, Agarwal PK, McDaniel MA, McDermott KB (2011) Test-enhanced learning in the classroom: long-term improvements from quizzing. J Exp Psychol Appl 17:382–395. doi: 10.1037/a0026252

69. RoedigerIII HL, Karpicke JD (2006) Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. Psychol Sci 17:249–255. doi: 10.1111/j.1467-9280.2006.01693.x

70. Rosé CP, McLaughlin EA, Liu R, Koedinger KR (2019) Explanatory learner models: Why machine learning (alone) is not the answer. Br J Educ Technol 50:2943–2958. doi: 10.1111/bjet.12858

71. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–215

72. Rudin C, Wang C, Coker B (2020) The age of secrecy and unfairness in recidivism prediction. Harv Data Sci Rev 2:1

73. Schmidt RM (2019) Recurrent Neural Networks (RNNs): A gentle Introduction and Overview

74. Semenova L, Rudin C (2019) A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. ArXiv

75. Shen S, Liu Q, Huang Z, Zheng Y, Yin M, Wang M, Chen E (2024) A survey of knowledge tracing: Models, variants, and applications. IEEE Trans Learn Technol

76. Shi Y, Schmucker R, Chi M, Barnes T, Price T (2023) KC-Finder: Automated Knowledge Component Discovery for Programming Problems. doi: 10.5281/ZENODO.8115671

77. Sonkar S, Ni K, Tran Lu L, Kincaid K, Hutchinson JS, Baraniuk RG (2024) Automated Long Answer Grading with RiceChem Dataset. In: Artificial Intelligence in Education: 25th International Conference, AIED 2024, Recife, Brazil, July 8–12, 2024, Proceedings, Part I. Springer-Verlag, Berlin, Heidelberg, pp 163–176

78. Stamper JC, Koedinger KR, Baker RSJ d., Skogsholm A, Leber B, Demi S, Yu S, Spencer D (2011) DataShop: A Data Repository and Analysis Service for the

Learning Science Community (Interactive Event). In: Biswas G, Bull S, Kay J, Mitrovic A (eds) Artificial Intelligence in Education. Springer, Berlin, Heidelberg, pp 628–628

79. Tsabari S, Segal A, Gal K (2023) Predicting Bug Fix Time in Students' Programming with Deep Language Models. doi: 10.5281/ZENODO.8115733

80. Tsutsumi E, Nishio T, Ueno M (2024) Deep-IRT with a Temporal Convolutional Network for Reflecting Students' Long-Term History of Ability Data. In: Olney AM, Chounta I-A, Liu Z, Santos OC, Bittencourt II (eds) Artificial Intelligence in Education. Springer Nature Switzerland, Cham, pp 250–264

81. Tuerlinckx F, Wang W-C (2004) Models for polytomous data. In: De Boeck P, Wilson M (eds) Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. Springer, New York, NY, pp 75–109

82. Tulis M, Steuer G, Dresel M (2016) Learning from errors: A model of individual processes. Frontline Learn Res 4:12–26. doi: 10.14786/flr.v4i2.168

83. Verguts T (2012) Computational Models of Human Learning. In: Seel NM (ed) Encyclopedia of the Sciences of Learning. Springer US, Boston, MA, pp 707–710

84. Wang X, Zheng Z, Zhu J, Yu W (2023) What is wrong with deep knowledge tracing? Attention-based knowledge tracing. Appl Intell 53:2850–2861

85. Weitekamp D, MacLellan C, Harpstead E, Koedinger K (2021) Decomposed Inductive Procedure Learning

86. Weitekamp D, Ye Z, Rachatasumrit N, Harpstead E, Koedinger K (2020) Investigating Differential Error Types Between Human and Simulated Learners. In: Bittencourt II, Cukurova M, Muldner K, Luckin R, Millán E (eds) Artificial Intelligence in Education. Springer International Publishing, Cham, pp 586–597

87. Wexler R (2017) When a computer program keeps you in jail. N Y Times 13:1

88. Yu H, Allessio DA, Rebelsky W, Murray T, Magee JJ, Arroyo I, Woolf BP, Bargal SA, Betke M (2024) Affect Behavior Prediction: Using Transformers and Timing Information to Make Early Predictions of Student Exercise Outcome. In: Olney AM, Chounta I-A, Liu Z, Santos OC, Bittencourt II (eds) Artificial Intelligence in Education. Springer Nature Switzerland, Cham, pp 194–208

89. Yudelson M, Pavlik PI, Koedinger KR (2011) User Modeling – A Notoriously Black Art. In: Konstan JA, Conejo R, Marzo JL, Oliver N (eds) User Modeling, Adaption and Personalization. Springer, Berlin, Heidelberg, pp 317–328

90. Yunsung Kim, Jadon Geathers, Chris Piech (2024) Grading and Clustering Student Programs That Produce Probabilistic Output. doi: 10.5281/ZENODO.12729772

91. Zambrano AF, Nasiar N, Ocumpaugh J, Goslen A, Zhang J, Rowe J, Esiason J, Vandenberg J, Hutt S (2024) Says Who? How different ground truth measures of emotion impact student affective modeling. pp 211–223

92. Zhang Q, Maclellan C (2021) Going Online: A simulated student approach for evaluating knowledge tracing in the context of mastery learning

93. Zhao Y, Qi Z, Do ST, Grossi J, Kang JH, Weiss GM (2024) Predicting GRE Scores from Application Materials in Test-Optional Admissions. pp 30–39

94. (2016) False Positives, False Negatives, and False Analyses: A Rejoinder to “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks.” https://www.uscourts.gov/about-federal-courts/probation-and-pretrial-services/federal-probation-journal/2016/09/false-positives-false-negatives-and-false-analyses-a-rejoinder-machine-bias-theres-software-used.

95. A computer program used for bail and sentencing decisions was labeled biased against blacks. It’s actually not that clear. - The Washington Post. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

96. The Bayesian information criterion: background, derivation, and applications - Neath - 2012 - WIREs Computational Statistics - Wiley Online Library. https://wires.onlinelibrary.wiley.com/doi/10.1002/wics.199. Accessed 14 May 2025

97. Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping | Science. https://www.science.org/doi/10.1126/science.1199327. Accessed 14 May 2025

98. (PDF) The Apprentice Learner Architecture: Closing the loop between learning theory and educational data. In: ResearchGate

99. Decomposed Inductive Procedure Learning: Learning Academic Tasks with Human-Like Data Efficiency | Proceedings of the AAAI Symposium Series. https://ojs.aaai.org/index.php/AAAI-SS/article/view/31289.