Thesis Proposal

Meaningful Models: Unlocking Insights Through Model Interpretations in Educational Data Mining

Napol Rachatasumrit

July, 2024

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

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: models that prioritize prediction accuracy often fail to provide scientifically or practically meaningful interpretations and explanations. Interpretations and explanations are crucial for scientific insight and are useful for practical applications, especially from the human-computer interaction perspective. For example, Deep Knowledge Tracing (DKT) 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, Additive Factor Model (AFM) 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). In this thesis, I argue for a claim that interpretations and explanations are what we need and not interpretable or explainable models that are not interpreted or explained, especially in the context of EDM. I aim to develop inherently interpretable or "meaningful" models that transcend post-hoc explanations of black-box models. Specifically, the variables and parameters of these meaningful models are associated with meaningful latent variables.

I make this argument with several examples of scenarios where the existing mechanisms or models are insufficient to produce meaningful interpretations and suggest strategies to investigate and fix them. For example, Performance Factor Analysis (PFA) has been demonstrated to outperform AFM, but we demonstrated that PFA parameters are confounded, which resulted in ambiguous interpretations. We then proposed an improved model that not only de-confound the parameters but also presented meaningful interpretations that lead to insights on the real-student datasets.

In my thesis, leveraging my experience from past projects, I propose generalized strategies for developing meaningful models and apply them to develop a model to capture spacing effect. Additionally, I will develop a recommender system to suggest an optimal study schedule based on the newly developed model to demonstrate the superiority of meaningful models, compared to black-box models, in practical applications. I will conduct in-vivo studies with middle school students in the biology domain to demonstrate the effectiveness of the system.

TABLE OF CONTENTS

ACRONYMS

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 [22]. 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: models that prioritize prediction accuracy often fail to provide scientifically or practically meaningful interpretations and explanations. While accurate prediction could be useful, the focus on prediction accuracy alone can overshadow the importance of understanding the underlying mechanisms and generalizable explanations driving these predictions. My goal is to develop inherently interpretable or "meaningful" models that go beyond post-hoc explanations of black-box models toward designing models that are inherently interpretable by providing not just better predictions but also estimations of parameters that provide an explanation of those predictions. Particularly, the parameters of these meaningful models correspond to latent variables, providing insights into the educational processes they represent.

Why are these inherently interpretable models important? Interpretations and explanations are crucial for scientific insight and are useful for practical applications, especially from the human-computer interaction perspective. Many existing models do provide either scientific insight or practical application. For example, Deep Knowledge Tracing (DKT) [17], a knowledge tracing model based on Recurrent Neural Network (RNN), 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 and there have been no scientific insights or practical applications resulting from it [21]. In contrast, Additive Factor Model (AFM) [3], 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) [13].

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 [6, 16, 20, 26, 28], while still providing valuable insights into the learning mechanisms and pedagogy [11, 13]. 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 [6, 14, 20]. 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 3.

In this thesis, I propose generalized strategies for developing meaningful models based on insights from my prior works. Then, I will demonstrate the utility of the proposed strategy by applying them to develop a new knowledge tracing model that effectively captures the spacing effect while maintaining the interpretable parameter estimates. Furthermore, to demonstrate the superiority of meaningful models, compared to black-box models, in practical applications, I will develop a recommender system to suggest an optimal study schedule based on the newly developed model. To complete my thesis, I propose conducting in-vivo studies with middle school students in the biology domain to demonstrate the effectiveness of the system.

This document is organized as follows:

In Chapter 2, I detail the background and related work. The first section briefly overviews the literature on models' explainability and interpretability, and the relevance of these arguments in the context of EDM. In the next section, I discuss the history and existing works on knowledge tracing models and models of human learning. In Chapter 3, I briefly describe examples of my related previous work in building meaningful models

and illustrate connections between these previous works and this thesis. Lastly, Chapter 4 describes the proposed work to complete my thesis. I discuss the proposed generalized strategies for developing meaningful models, and provide further detail on the new model, its application, and the proposed in-vivo study.

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 [27]. Similarly the field of EDM has prioritized prediction accuracy such that black-box models have been increasingly used [5, 7]. 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.

While some believe that developing methods to explain these black-box models can mitigate these issues [10], 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 [22]. 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 [24]. 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 examples of my previous work developing models that are interpretable by this expanded definition.

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 [12], 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) [4], which models student mastery as a latent variable in a simple Hidden Markov Model [2], has been widely used in the real-world ITSs and shown to be reasonably effective for mastery learning and problem selection [25]. 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) [15]. 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 [6, 20, 26, 28].

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 [6, 16, 20, 26].

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 [11, 13]. 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.

CHAPTER 3
PRIOR WORK IN BUILDING MEANINGFUL MODELS

3.1 Content Matters: A Computational Investigation into the Effectiveness of Retrieval Practice and Worked Examples

This section was adapted from my published work [18]:

> Rachatasumrit, N., Carvalho, P., Koedinger, K. (2023), Content Matters: A Computational Investigation into the Effectiveness of Retrieval Practice and Worked Examples. AIED 2023, Proceedings The 24th International Conference on Artificial Intelligence in Education

In this work we argue that artificial intelligence models of learning can contribute precise theory to explain surprising student learning phenomena. In some past studies of student learning, practice produces better learning than studying examples, whereas other studies show the opposite result. We reconcile and explain this apparent contradiction by suggesting that retrieval practice and example study involve different learning cognitive processes, memorization and induction, respectively, and that each process is optimal for learning different types of knowledge. We implement and test this theoretical explanation by extending an AI model of human cognition — the Apprentice Learner Architecture (AL) — to include both memory and induction processes and comparing the behavior of the simulated learners with and without a forgetting mechanism to the behavior of human participants in a laboratory study. We show that, compared to simulated learners without forgetting, the behavior of simulated learners with forgetting matches that of human participants better. Simulated learners with forgetting learn best using retrieval practice in situations that emphasize memorization (such as learning facts or simple associations), whereas studying examples improves learning in situations where there are multiple pieces of information available and induction and generalization are necessary (such as when learning skills or procedures).

This work is an example of a computational model that mainly focuses on the underlying mechanisms, which make them inherently interpretable. It allows us to make changes (e.g. adding a memory mechanism) that enhance its capability to model new

phenomena while maintaining its interpretability. From the experiment, we also gained insight into how inductive learning and memory mechanisms interact differently with different types of content (e.g. fact vs skill).

3.2 Good Fit Bad Policy: Why Fit Statistics are a Biased Measure of Knowledge Tracer Quality

This section was adapted from my published work [20]:

> Rachatasumrit, N., Weitecamp, D., Koedinger, K. (2024), Good Fit Bad Policy: Why Fit Statistics are a Biased Measure of Knowledge Tracer Quality. AIED 2024, Proceedings The 25th International Conference on Artificial Intelligence in Education.

Knowledge tracers are typically evaluated on the basis of the goodness-of-fit of their underlying student performance models. However, for the purposes of supporting mastery learning the true measure of a good knowledge tracer is not its goodness-of-fit, but the degree to which it optimally selects next problem items. In this context, a knowledge tracer should minimize under-practice to ensure students master learning materials and minimize over-practice to reduce wasted time. Prior work has suggested that fit-statistic-based measures of knowledge tracer quality may misrank the relative quality of knowledge tracers' item selection. In this work, we evaluate this claim by measuring over- and under-practice directly in synthetic data drawn from ground-truth learning curves. We conduct an experiment with 3 well-known student performance models: Performance Factor Analysis (PFA), BestLR, and Deep Knowledge Tracing (DKT), and find that in 43% of the synthetic datasets, the models with higher measures of overall predictive performance (e.g. AUC and MSE) were worse than a comparison model with a lower predictive performance at minimizing the number of predictions that would lead to over-practice and under-practice attempts. 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 bring into question the validity of traditional methods of knowledge tracer comparison.

In this work, we have demonstrated that black-box models like DKT do not always yield a better prediction performance compared to more interpretable models like PFA and BestLR, especially when we consider the metrics that are directly related to real-world applications like number of over-practices and under-practices.

3.3 Toward Improving Student Model Estimates through Assistance Scores in Principle and in Practice

This section was adapted from my published work [19]:

> Rachatasumrit, N., Koedinger, K.R. (2021), Toward Improving Student Model Estimates through Assistance Scores in Principle and in Practice. EDM 2021: Proceedings of the 14th International Conference on Educational Data Mining.

Student modeling is useful in educational research and technology development due to a capability to estimate latent student attributes. Widely used approaches, such as the Additive Factors Model (AFM), have shown satisfactory results, but they can only handle binary outcomes, which may yield potential information loss. In this work, we propose a new partial credit modeling approach, PC-AFM, to support multi-valued outcomes. We focus particularly on the amount of assistance, that is, the number of error feedback and hint messages a student needs to get a problem step correct. Because errors and hint requests may not only derive from student ability, but also from non-cognitive factors (e.g., students may game the system), we first test PC-AFM on synthetic data where non-cognitive factors source of variation is not present. We confirm that PC-AFM is indeed better than AFM in recovering the true student and knowledge component (KC) parameters and even predicts student error rates better than a model fit to error rates. We then apply the approach to six real-world datasets and find that PC-AFM outperforms AFM in reliable estimation of KC parameters and produces better generalization to new students, which requires better KC estimates. However, consistent with the hypothesis that student assistance behavior is driven by motivational or meta-cognitive factors beyond their ability, we found that PC-AFM was not better in reliable estimation of student parameters nor in generalization across items, which

requires accurate student estimates. We propose cross-measure cross-validation as a general method for comparing alternative measurement models for the same desired latent outcome.

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.

3.4 Beyond Accuracy: Embracing Meaningful Parameters in Educational Data Mining

This section was adapted from my published work:

> Rachatasumrit, N., Carvalho, P.F., Koedinger, K.R. (2024), Beyond Accuracy: Embracing Meaningful Parameters in Educational Data Mining. EDM 2024: Proceedings of the 17th International Conference on Educational Data Mining.

What does it mean for a model to be a better model? One conceptualization, indeed a common one in Educational Data Mining, is that a better model is the one that fits the data better, that is, higher prediction accuracy. However, oftentimes, models that maximize prediction accuracy do not provide meaningful parameter estimates, making them less useful for building theory and practice. Here we argue that models that provide meaningful parameters are better models and, indeed, often also provide higher prediction accuracy. To illustrate our argument, we investigate the Performance Factor Analysis (PFA) model and the Additive Factors Model (AFM). PFA often has higher prediction accuracy than the AFM. However, PFA's parameter estimates are ambiguous and confounded. We propose more interpretable models (AFMh and PFAh) designed to address the confounded parameters and use synthetic data to demonstrate PFA's

parameter interpretability issues. The results from the experiment with 27 real-world datasets also support our claims and show that more meaningful models will also produce better predictions.

CHAPTER 4
PROPOSED WORK

In my research so far, I have shown that inherently interpretable or “meaningful” models based on cognitive models can be as capable as black-box models when it comes to predicting student learning outcomes. Meaningful models have an advantage in that they also lead to insights, both practical and scientific, that result in actual applications and use cases. For instance, while the Additive Factor Model (AFM) often falls short of DKT in prediction accuracy, its parameter estimates offer valuable interpretations (such as the slope representing the learning rate of knowledge components). These insights can drive new scientific discoveries (like enhanced cognitive models) and lead to practical applications (such as redesigning intelligent tutoring systems). In prior work, we have explored meaningful models that capture the non-binary nature of students’ learning outcomes, the impact of content types on learning, and student-KC interactions, and demonstrated the insights that we acquired from those models. Based on this work, I propose generalized strategies for developing meaningful models:

> S1: When you identify a confounded latent variable in the predictive model, formulate a new observable or computed variable/measure? and a new associated latent variable to help remove the confound.
>
> S2: When a complex less meaningful model predicts better than a simpler more meaningful model, try to find out under what circumstances it is better. Use those circumstances to hypothesize and test new observed, computed, and/or latent variables.
>
> S3: When you discover that a complex model (e.g., a deep learning model or LLM) is identifying an indicator variable that is masking a correlated causal variable, create a model that prefers the causal variable.

As an example, my prior work on PFAh illustrates the application of these strategies. In that work, I demonstrated that PFA (a complex model) predicts better than AFM (a

simpler model) particularly when there are differences in learning rate between successful and failed attempts or strong student-KC interactions (S2). In the work, the success and failure slopes in PFA are also shown to be confounded by differences in learning rate and student-KC interactions (S1). The new computed variable based on the ratio of the history of success attempts was then proposed to help remove the confound, which led to the new model that was more interpretable.

In the proposed work, I will develop a new meaningful model in a new domain utilizing these strategies. In particular, the learning principle and data that I will focus on relate to the spacing effect. Spacing is one of the most important attributes that could affect learning, and its impact has been studied to a great extent; however, there is no existing quantitative model that focuses on the parameters interpretability. Additionally, unlike previous work where I have only discussed potential use cases of the acquired insights from the models, I will close the loop by developing an application based on the new model and conduct an experimental comparison that tests whether an application built based on an meaningful model is more effective than one based on a black-box model. In this proposed work, I will

1. Develop a knowledge tracing model that captures spacing effects while preserving meaningful interpretation of their parameter estimates.
2. Build a system based on the knowledge tracing model that can suggest optimal study schedule and evaluate it with real students.
3. Conduct an in vivo study with Podsie to evaluate the system from (2) and compare the system against systems that are based on black-box models. (See details in Evaluation)

4.1 Data

During the model development, I will use a combination of real-student data and simulated data. The real student data will consist of existing data from DataShop and newly collected data with Podsie, a personalized learning tool, from middle school STEM students (Biology, Physics, Math, and Chemistry classes). The simulated data will be generated, based on a combination of ACT-R memory model, which can be used to

calculate the retrieval probabilities of information based on its activation levels, and some existing knowledge tracing models, varying among different content types, spacing, KC's difficulty, and students' ability. Afterwards, we will deploy the knowledge tracing model in the in-vivo study by integrating it in Podsie to suggest the optimized study schedule to students and evaluate the student's learning outcome improvement from following the suggested schedule via Podsie.

4.2 Developing a meaningful model of Spacing

The main challenge for this new model is to successfully employ the strategies to identify and add new, maybe computed, variables that are capable of capturing spacing effects and include them in the linear models. This knowledge tracer would add a model forgetting and spacing effect to the existing learning growth component in the logistic regression AFM family of models (AFM, iAFM, or PFAh). Examples of potential variables are a geometric mean of the spaces between prior attempts or a variable based on the average distance between optimal and actual spacing, where we computed optimal spacing from the number of prior opportunities and duration from last opportunity to predicted one. These variables will allow us to capture a sequence of spaces in a single variable, but the downside is that it ignores the order of the prior attempts which is an important part of the spacing effect when an expanding strategy is used. During the development, I will come up with a list of potential variables and experiment with both simulated data and real-student data to evaluate the pros and cons for each of them [26].

4.3 Demonstrate an Application that uses the Meaningful Model works better than one using a Complex Black-box model

To demonstrate that meaningful models are more effective in practical applications, I will leverage the newly developed knowledge tracing model to build a recommender system that suggests an optimal practice schedule for a student-KC pair given a desired retention interval, the time of the first and last practice, and the number of repetitions. I

will investigate two potential solutions for the proposed system. Firstly, I will attempt to derive a closed-form solution from the parameters learning from fitting the data in the model. This approach will only be possible with inherently interpretable models that allow us to attain meaningful parameters, in contrast to black-box models where there is no meaningful parameter to utilize. Another solution is to develop a policy for optimal practice schedule using reinforcement learning approaches and leverage the knowledge tracing model in a reward or Q-value function.

4.4 Timeline of Completion

My goal is to complete the dissertation by May 2025. This gives me enough time to develop the new knowledge tracing models and the optimal schedule recommender, and conduct the in-vivo studies with Podsie in the Spring 2025. My proposed schedule is shown below:

- July 22, 2024: Thesis proposal
- August 2024 - October 2024: Model and application development
- Late Fall 2024 / Early Spring 2025: In-vivo study with Podsie
- April 2025 - May 2025: Thesis writing & Defense

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. 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
3. 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
4. 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
5. 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
6. 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
7. Gillani N, Eynon R, Chiabaut C, Finkel K (2023) Unpacking the “Black Box” of AI in education. Educ Technol Soc 26:99–111
8. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36. doi: 10.1148/radiology.143.1.7063747
9. Jaspen N (1968) Applied Regression Analysis
10. 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
11. 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

12. 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
13. 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
14. Mandalapu V, Gong J, Chen L (2021) Do we need to go Deep? Knowledge Tracing with Big Data
15. 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
16. Pavlik PI, Eglington LG Automated Search Improves Logistic Knowledge Tracing, Surpassing Deep Learning in Accuracy and Explainability
17. 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.
18. 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
19. Rachatasumrit N, Koedinger KR (2021) Toward Improving Student Model Estimates through Assistance Scores in Principle and in Practice. International Educational Data Mining Society
20. 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
21. 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

22. 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
23. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 461–464
24. Semenova L, Rudin C (2019) A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. ArXiv
25. 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
26. 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
27. Wexler R (2017) When a computer program keeps you in jail. N Y Times 13:1
28. Zhang Q, Maclellan C (2021) Going Online: A simulated student approach for evaluating knowledge tracing in the context of mastery learning