Dynamic Bayesian Networks for Student Modeling

Tanja Käser, Severin Klingler, Alexander G. Schwing, and Markus Gross

Abstract—Intelligent tutoring systems adapt the curriculum to the needs of the individual student. Therefore, an accurate representation and prediction of student knowledge is essential. Bayesian Knowledge Tracing (BKT) is a popular approach for student modeling. The structure of BKT models, however, makes it impossible to represent the hierarchy and relationships between the different skills of a learning domain. Dynamic Bayesian networks (DBN) on the other hand are able to represent multiple skills jointly within one model. In this work, we suggest the use of DBNs for student modeling. We introduce a constrained optimization algorithm for parameter learning of such models. We extensively evaluate and interpret the prediction accuracy of our approach on five large-scale data sets of different learning domains such as mathematics, spelling learning, and physics. We furthermore provide comparisons to previous student modeling approaches and analyze the influence of the different student modeling techniques on instructional policies. We demonstrate that our approach outperforms previous techniques in prediction accuracy on unseen data across all learning domains and yields meaningful instructional policies.

Index Terms—Bayesian networks, parameter learning, constrained optimization, error measures, instructional policies

1 Introduction

Modeling and predicting student knowledge is a fundamental part of an intelligent tutoring system (ITS). In these systems, the selection of challenges to work on next is based on the estimation and prediction of the student's current knowledge by the student model. The prediction accuracy and behavior of the student model directly influence the instructional policies [1] of the system and hence also the quality of teaching decisions. Therefore, an accurate student model is essential for individualization, i.e., the adaptation of the learning content and the difficulty level to the individual student. A large body of research was conducted in terms of *modeling*, i.e., the construction of accurate student models, and in terms of *assessment*, i.e., the definition of appropriate error measures for assessing prediction accuracy of student models.

Research in *modeling* covers a wide range of approaches for student modeling. Common techniques include logistic regression models and probabilistic models. Two of the most popular algorithms for estimating student knowledge are performance factors analysis (PFA) [2] and Bayesian Knowledge Tracing (BKT) as presented by Corbett and Anderson [3].

- *T. Käser is with the Graduate School of Education, Stanford University, Stanford, CA 94305. E-mail: tkaeser@stanford.edu.*
- *S. Klingler and M. Gross are with the Department of Computer Science, ETH Zurich, 8006 Zurich, Switzerland. E-mail: {kseverin, grossm}@inf.ethz.ch.*
- *A.G. Schwing is with the Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, IL 61820-6983. E-mail: aschwing@illinois.edu.*

Manuscript received 21 Mar. 2016; revised 7 Dec. 2016; accepted 16 Mar. 2017. Date of publication 29 Mar. 2017; date of current version 13 Dec. 2017. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TLT.2017.2689017

Traditional BKT and improved versions have been successfully used in different tutoring systems [4], [5]. An important task when using BKT is parameter learning. BKT models have been fit using brute-force grid search [6], gradient descent [7] or expectation maximization [8]. A large body of research was conducted to construct BKT models with higher prediction accuracy. Previous work includes clustering approaches [9] or individualization techniques, such as learning student- and skill-specific parameters [7], [10], [11] or modeling the parameters per school class [12].

Exhibiting a tree structure (directed acyclic variable dependencies with a single root variable), BKT allows for efficient parameter learning and accurate inference. However, tree-like models lack the ability to represent the hierarchy and relationships between the different skills of a learning domain. Employing dynamic Bayesian network models (DBN) has the potential to increase the representational power of the student model by explicitly modeling prerequisite skill hierarchies and hence further improve prediction accuracy. Furthermore, hierarchical models allow to incorporate expert knowledge and implicitly define the curriculum of the system. Indeed, hierarchical models have been successfully applied in ITS. Dynamic learning maps structure the sequence of topics hierarchically to reach certain thresholds in post-tests [13]. In cognitive item response theory, the attribute hierarchy method models hierarchically ordered attributes (competencies) required to correctly solve test problems [14]. Bayesian networks [15], [16], [17] have been applied to model skill dependencies at a single point in time.

DBNs, on the other hand, have been used to model and predict students' performance [18], [19], engagement states [20], [21], and goals [18]. DBNs are also employed in user modeling [22]. In cognitive sciences, DBNs are applied to model human learning [23] and understanding [24]. Despite

their beneficial properties to represent knowledge, DBNs have received less attention in modeling skill hierarchies as the joint representation of multiple skills over time in one network imposes challenges for learning and inference.

Lately, research in *assessment* of student models has gained increasing interest. Traditionally, the prediction accuracy of student models has been assessed using performance measures such as the root mean squared error (RMSE) [7], [12], [25], [26], the area under the ROC curve (AUC) [6], [27], [28], [29], [30], the mean absolute error (MAE) [31], [32] and the log-likelihood (LL) [2], [33]. Recently, the use of a single number as a performance measure has been questioned [34] and the different measures have been analyzed and discussed [34], [35]. Furthermore, new measures [34] and frameworks [36], [37] capturing model behavior in more detail have been introduced. Recent work has also analyzed the impact of the prediction accuracy on instructional policies [1].

In this paper, we contribute to both the research area of *modeling* and the research area of *assessment*: We present an efficient method for modeling prerequisite skill hierarchies and extensively evaluate our approach based on the latest research conducted in assessment of student models. We demonstrate how the topologies and relationships between different skills of a learning domain can be modeled using DBNs and show the advantages of using pre-requisite hierarchies for student modeling. Recently, [38] showed that a constrained latent structured prediction approach to parameter learning for DBNs yields accurate and interpretable models. Furthermore, DBNs modeling prerequisite skill hierarchies outperform traditional BKT regarding prediction accuracy [39]. We extend [39] by extensively evaluating the behavior of our hierarchical skill models regarding prediction accuracy and instructional policies and providing comparisons to the popular traditional approaches to student modeling. We define domain-specific DBNs modeling skill hierarchies for five large-scale data sets from different learning domains, containing up to 7,000 students. To analyze the generalization capabilities of our approach, we have selected data sets generated from different tutoring systems addressing different fields of education such as mathematics, spelling learning and physics. Furthermore, students' age ranges from elementary school to university level. Our results show that even simple skill hierarchies lead to significant improvements in prediction accuracy over BKT and PFA across all learning domains. Furthermore, the hierarchical skill models lead to meaningful instructional policies. By using the same constraints and parametrization for all our experiments, we also demonstrate that basic assumptions about learning hold across different learning domains.

2 Background: Student Models

In this section, we provide an overview of popular approaches for student modeling, which we later use for comparison in the experimental evaluation of our method.

2.1 Latent Factors Models

Student models applying logistic regression are based on the idea that the probability of a correct response to a task can be represented by a mathematical function of student and skill parameters. In such models, the binary task outcomes (correct/wrong) of the students follow a Bernoulli distribution, i.e., a binomial distribution with $n = 1$. Letting $y_{st} \in \{0, 1\}$ denote the response of student s to task t, we obtain $y_{st} \sim binomial(1, p_{st})$.

Additive Factors Model (AFM). The AFM [40], [41] is a logistic regression model, which models the probability p_{st} of solving a task t correctly as a function of three parameters, resulting in

$$p_{st} = (1 + \exp(-(\theta_s + \sum q_{kt}(\beta_k + \gamma_k \cdot T_{sk}))))^{-1}. \quad (1)$$

Hereby the random effect $\theta_s \sim \mathcal{N}(0, \sigma_\theta^2)$ denotes the student proficiency and the fixed effects β_k and γ_k denote the difficulty and the learning rate for skill k, respectively. T_{sk} is the number of practice opportunities student s has seen for skill k and q_{kt} is 1, if task t uses skill k, and 0 otherwise.

Performance Factors Analysis (PFA). The PFA [2] is an extension of the AFM, with the following definition of the linear component

$$p_{st} = (1 + \exp(-(\theta_s + \sum q_{kt}(\beta_k + \gamma_k \cdot S_{sk} + \rho_k \cdot F_{sk}))))^{-1}. \quad (2)$$

In contrast to AFM, PFA differentiates between correct and incorrect practice opportunities: S_{sk} denotes the number of correctly solved tasks for student s at skill k, while F_{sk} denotes the number of failures for student s at skill k. The fixed effects γ_k and ρ_k therefore denote the learning rates for successes and failures, respectively.

2.2 Bayesian Knowledge Tracing

BKT models are a special case of DBNs [42] or more specifically of Hidden Markov Models (HMM), consisting of observed and latent variables. Latent variables represent student knowledge about one specific skill and are assumed to be binary, i.e., a skill can either be mastered by the student or not. The observed variables are also binary: they represent student answers (correct/incorrect) to questions associated with a specific skill and depend directly on the latent variables. The state of the latent variables is inferred based on these observations.

There are two types of parameters in an HMM: transition probabilities and emission probabilities. In BKT, the emission probabilities are defined by the slip probability p_S of making a mistake when applying a known skill and the guess probability p_G of correctly applying an unknown skill. The transition probabilities are described by the probability p_L of a skill transitioning from the unknown to the known state, while p_F is the probability of forgetting a previously known skill. In BKT, p_F is assumed to equal zero. The last parameter required to describe the BKT model is the initial probability p_0 of knowing a skill a-priori.

Employing one BKT model per skill, the learning task amounts to estimating the parameters given some observations. More specifically, given a sequence of observations $\mathbf{y_m} = (y_{m,1}, \ldots, y_{m,T})$ with $y_{m,t} \in \{0, 1\}$ and time $t \in \{1, \ldots, T\}$ for the mth student with $m \in \{1, \ldots, M\}$, what are the parameters $\theta = \{p_0, p_L, p_F, p_S, p_G\}$ that maximize the likelihood $\prod_m p(\mathbf{y_m} \mid \theta)$ of the available data. In BKT, this task is

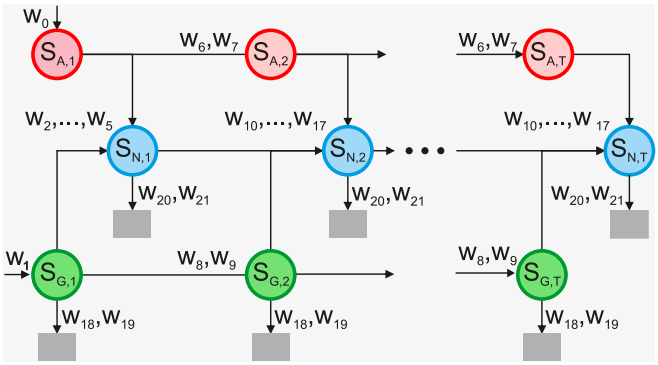

Fig. 1. Structure of the graphical model for a DBN unrolled over T time steps. Circular nodes represent skills (latent binary variables), while the rectangles represent the observable student answers to tasks associated with those skills. The illustrated DBN models the hierarchical relationships between the three skills S_A, S_G, and S_N. The index numbers denote the respective time step, i.e., $S_{A,1}$ represents the state of skill S_A at time $t = 1$.

commonly solved using brute-force grid search [6], gradient descent [7] or expectation maximization [8].

3 Methods: Dynamic Bayesian Networks

DBNs have the potential to increase the representational power of the student model and therefore further increase prediction accuracy. Subsequently, we first give an introduction to DBNs using the traditional probabilistic notation, before introducing our log-linear formulation and the corresponding optimization. Finally, we also present a possible specification of a simple DBN.

When employing DBNs, we consider the different skills of a learning domain jointly within a single model. Student knowledge is again represented using binary latent variables (one per skill), which are updated based on observations associated with the skill under investigation. However, we now also model the dependencies between the different skills, e.g., two skills S_A and S_B are conditionally dependent if S_A is a prerequisite for mastering S_B.

To illustrate this concept, consider Fig. 1 which shows an example DBN unrolled over T time steps. The circular nodes represent binary skill variables: similar to BKT, a skill can be either mastered or not. The example network consists of three skills: S_A, S_G, and S_N. The rectangular nodes represent observable variables, i.e., binary student answers (correct/incorrect) to tasks associated with the respective skill. The index numbers indicate the time step.

3.1 Probabilistic Notation

The learning task of a DBN model is described as follows: let the set of N variables of the model be denoted by $X = \{X_i \,|\, i \in \{1, \ldots, N\}\}$. The set of variables X contains all skill nodes S as well as all observation nodes O of the model. In addition, let $\mathcal{H}$ denote the domain of the unobserved variables, i.e., missing student answers and the binary skill variables, while $\mathcal{Y}$ refers to the observed space, disjoint from the latent space $\mathcal{H}$. Let us consider an example student m for the model in Fig. 1 solving a task associated with skill $S_{G,1}$ in the first time step correctly, i.e., $O_{G,1} = 1$. The set of observed variables $\mathbf{y_m}$ for the first time step consists of the student's answer, i.e., $\mathbf{y_m} = \{o_{G,1}\}$. The set of hidden variables $\mathbf{h_m}$ for the first time step contains all other variables, i.e., $\mathbf{h_m} = \{S_{A,1}, S_{N,1}, S_{G,1}, O_{N,1}\}$. During learning, we are interested in finding the parameters θ that maximize the likelihood of the observed data $\bigcup_m \mathbf{y_m}$ with $\mathbf{y_m} = (y_{m,1}, \ldots, y_{m,T})$ representing a sequence of T binary answers from the mth student, while $\mathbf{h_m}$ denotes the hidden variables for student m. The log-likelihood of a DBN [43] is then given by

$$L(\theta) = \sum_m \ln\left(\sum_{\mathbf{h_m}} p(\mathbf{y_m}, \mathbf{h_m} \,|\, \theta)\right), \tag{3}$$

where we marginalize over the states of the hidden variables $\mathbf{h_m}$ for student m. The joint probability $p(\mathbf{y_m}, \mathbf{h_m} \,|\, \theta)$ of the model for student m is defined via

$$\begin{aligned} p(\mathbf{y_m}, \mathbf{h_m} \,|\, \theta) &= \prod_i p(X_{m,i} = x_{m,i} \,|\, pa(X_{m,i}) = \mathbf{x_{m,pa(X_{m,i})}}) \\ &= \prod_i p_{ij_{m,i}\mathbf{k_{m,i}}}, \end{aligned} \tag{4}$$

where $X_{m,i}$ is the ith variable in the model for student m and $pa(X_{m,i})$ are the parents of $X_{m,i}$, while $x_{m,i}$ and $\mathbf{x_{m,pa(X_{m,i})}}$ are the realizations of the random variables $X_{m,i}$ and $pa(X_{m,i})$, i.e., the states assigned to $X_{m,i}$ and $pa(X_{m,i})$ given by $(\mathbf{y_m}, \mathbf{h_m})$. Furthermore, we let $j_{m,i} := x_{m,i}$ and $\mathbf{k_{m,i}} := \mathbf{x_{m,pa(X_{m,i})}}$ to simplify the notation. Therefore, $p_{ij_{m,i}\mathbf{k_{m,i}}}$ denotes exactly one entry in the conditional probability table (CPT) of $X_{m,i}$.

3.2 Log-Linear Models

The log-likelihood of a DBN can alternatively be formulated using a log-linear model. This formulation is flexible and predominantly used in literature [44], [45]. Therefore, we reformulate the learning task in the following. Let $\boldsymbol{\phi} : \mathcal{Y} \times \mathcal{H} \to \mathbb{R}^F$ denote a mapping from the latent space $\mathcal{H}$ and the observed space $\mathcal{Y}$ to an F-dimensional feature vector. The log-likelihood from Eq. (3) can then be reformulated to

$$L(\mathbf{w}) = \sum_m \ln\left(\sum_{\mathbf{h_m}} \exp\left(\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{y_m}, \mathbf{h_m}) - \ln(Z)\right)\right), \tag{5}$$

where Z is a normalizing constant and $\mathbf{w}$ denotes the weights of the model. Next, we show that this log-linear formulation of the log-likelihood is equivalent to the traditional notation. Comparing Eqs. (5) to (3), it follows that

$$\begin{aligned} \prod_i p_{ij_{m,i}\mathbf{k_{m,i}}} &= \frac{1}{Z} \exp \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{y_m}, \mathbf{h_m}) \\ &= \frac{1}{Z} \exp \sum_i w_i^\top \boldsymbol{\phi}_i(\mathbf{y_m}, \mathbf{h_m}), \end{aligned} \tag{6}$$

and therefore

$$\forall i, j, \mathbf{k} : p_{ij\mathbf{k}} = \frac{1}{Z} \exp w_i^\top \boldsymbol{\phi}_i(\mathbf{x}), \tag{7}$$

where $\mathbf{x}$ are the realizations of all random variables in X with $j \in \mathbf{x}$ and $\mathbf{k} \subset \mathbf{x}$. A feature vector $\boldsymbol{\phi}$ and weights $\mathbf{w}$ that fulfill Eq. (7) can be specified as follows: consider the CPT describing the relationship between a node X_A and its $n-1$ parent nodes $pa(X_A)$. The CPT for these n nodes contains 2^n

entries. Let $\mathbf{k} \in \{0,1\}^{n-1}$ denote one possible assignment of states to the parent nodes $pa(X_A)$. We can therefore define $p(X_A = 1 \,|\, pa(X_A) = \mathbf{k}) = 1 - p(X_A = 0 \;|\, pa(X_A) = \mathbf{k}) = 1 - p_{A,0,\mathbf{k}}$. To continue, let $p_{A,x_A,\mathbf{k}} = \frac{1}{Z} \exp w_{A,\mathbf{k}} \;\; (1 - 2x_A) = \exp w_{A,\mathbf{k}}(1 - 2x_A)/(\exp w_{A,\mathbf{k}} + \exp(-w_{A,\mathbf{k}}))$, which leads to the feature function $\phi_A(x) = 1 - 2x_A$. We therefore obtain the joint distribution as a product of the exponential terms which translates to a weighted linear combination of feature vector entries in the exponent and thus fulfills Eq. (7). From this formulation also follows that we need 2^{n-1} parameters to specify a CPT including n skills.

3.3 Optimization

Due to the loopy structure, the learning task for DBNs is computationally more expensive than the one for HMMs. However, [45] showed that some approximations admit efficient parameter learning. Note that interpretability of the parameters is not ensured, since guarantees exist only for converging to a local optimum, i.e., the approximation might result in degenerate models with for example a probability of guessing $p_G > 0.5$. Therefore, [38] extended the approach presented by [45] to include constraints on parameters and demonstrated that the constrained optimization increases prediction accuracy on unseen data while guaranteeing interpretable models. Using the log-linear formulation, the algorithm presented in [38] can be directly applied to learn the parameters of a DBN model.

3.4 Specification

Next, we explain the specification of a simple DBN. The example DBN illustrated in Fig. 1 consists of three skills: S_A, S_G, and S_N. Two of the skills (S_G and S_N) have tasks associated with them, while the third skill S_A cannot be observed. Similarly to BKT, we can interpret the parameters of a DBN in terms of a learning context. To specify the CPTs of the example DBN, we employ $F = 22$ weights that can be associated with a parameter set θ. We subsequently use $\simeq$ to denote proportionality in the log domain; i.e., $w \simeq p$ is equivalent to $w \propto \exp p$. Let O_N denote the task associated with skill S_N. Then the parameters $w_{20} \simeq p(O_N = 0 \,|\, S_N = 0) = 1 - p_G$ and $w_{21} \simeq p(O_N = 0 \,|\, S_N = 1) = p_S$ represent the guess and slip probabilities. Similarly, w_{18} and w_{19} are associated with p_G and p_S as evident from Fig. 1. Furthermore, parameters $w_6 \simeq p(S_{A,t} = 0 \,|\, S_{A,t-1} = 0) = 1 - p_L$ and $w_7 \simeq p(S_{A,t} = 0 \,|\, S_{A,t-1} = 1) = p_F$ are associated with learning and forgetting; the same holds true for w_8 and w_9.

Skills S_A and S_G are prerequisites for knowing skill S_N, i.e., the probability that skill S_N is mastered in time step t depends not only on the state of skill S_N in the previous time step, but also on the states of S_A and S_G in the current time step. Therefore $w_{10} \simeq p(S_{N,t} = 0 \,|\, S_{N,t-1} = 0, S_{A,t} = 0, S_{G,t} = 0) = 1 - p_{L0}$, where p_{L0} denotes the probability that the student learns S_N despite not knowing S_A and S_G. Also, $w_{17} \simeq p(S_{N,t} = 0 \,|\, S_{N,t-1} = 1, S_{A,t} = 1, S_{G,t} = 1) = p_{F1}$, the probability of forgetting a previously learnt skill. Furthermore, we set $w_l \simeq 1 - p_{LM}$ if $l \in \{11, 12, 13\}$ and $w_l \simeq 1 - p_{FM}$ if $l \in \{14, 15, 16\}$, where p_{LM} denotes the probability that the student learns S_N given that he knows at least one of the precursor skills of S_N. Moreover, p_{FM} is the probability that the student forgets the previously known skill S_N, when either S_A or S_G or none of them are known.

Finally, the parameters w_l with $l \in \{2, 3, 4, 5\}$ describe the dependencies between the different skills. We let $w_l \simeq 1 - p_{P0}$, if $l \in \{2, 3, 4\}$ and $w_5 \simeq p_{P1}$, where p_{P0} is the probability of knowing a skill despite having mastered only part of the prerequisite skills and p_{P1} denotes the probability of failing a skill given that all precursor skills have been mastered already. Moreover, we refer to the probability of knowing a skill a-priori via p_0. Note that w_0 and w_1 are associated with p_0. The example DBN can therefore be described by the parameter set $\theta = \{p_0, p_G, p_L, p_F, p_{L0}, p_{F1}, p_{LM}, p_{FM}, p_{P0}, p_{P1}\}$. Importantly, the method proposed in this work is independent of the exact parametrization used. The parametrization introduced here serves as an example only and could be easily extended.

4 Experimental Evaluation

We show the benefits of DBN models with higher representational power on five data sets from various learning domains. The data sets were collected with different tutoring systems and contain data from elementary school students up to university students. With the extensive experimental evaluation, we aim at answering the following three research questions. *1) What are the benefits in terms of prediction accuracy of DBNs representing skill hierarchies compared to traditional student modeling approaches assuming independence between the different skills? 2) Where do the differences in prediction accuracy between the different student models come from? 3) How do the different student modeling techniques influence instructional policies?*

4.1 Experimental Setup

Evaluation of the models was performed using a training-test setting: For all the different models, we learned the parameters on a training set and evaluated their performance on a test set. All evaluation measures were computed using student-stratified 10-fold cross validation.

Fitting the BKT models was done using [7], applying skill-specific parameters and using gradient descent for optimization. As described in [7], we set the forget probability p_F to 0, while p_S and p_G are bounded by 0.3. In the following, we will denote this constrained BKT version as BKT$_C$. Since parameter fitting for BKT models converges to different values depending on the initialization of the algorithm [6], [28], [46], we generated ten different BKT$_C$ models using different start values. We included the default values provided by [7] as well as nine other randomly generated sets of start values fulfilling the constraints. Subsequently, we selected the model with the best prediction accuracy in terms of root mean squared error on the test set for comparison. This selection process overestimates the performance of BKT and therefore provides an upper bound on the prediction accuracy of BKT.

The parameters of the latent factors models were trained using the `lme4` package of `R`. AFM and PFA require a student parameter (the student proficiency θ_p): For the unseen students in the test sets, we set θ_p to the mean of the trained student parameters.

We used constrained latent structured prediction [38] to learn the parameters of the DBNs. All models were parametrized according to Section 3.4 and we imposed the constraints described in the following on the parameter set θ of

the different models to ensure interpretable parameters. For our first constraint set $\mathcal{C}_1$, we let $p_D \leq 0.3$ for $D \in \{G, S, L, F, L0, F1\}$ to ensure that parameters associated with guessing, slipping, learning and forgetting remained plausible. The constraints on θ can be directly turned into constraints on $\mathbf{w}$. For the example DBN (Fig. 1), the constraints translate into the following linear constraints on the weights for $\mathcal{C}_1$: $w_i \geq 0.4236$, if $i \in \{6, 8, 10, 18, 20\}$ and $w_i \leq -0.4236$, if $i \in \{7, 9, 17, 19, 21\}$. For the second constraint set $\mathcal{C}_2$, we augmented $\mathcal{C}_1$ by limiting $p_D \leq 0.3$ if $D \in \{LM, FM, P0, P1\}$, yielding $w_i \geq 0.4236$, if $i \in \{2, 3, 4, 11, 12, 13\}$ and $w_i \leq -0.4236$, if $i \in \{5, 14, 15, 16\}$ for the example DBN (Fig. 1). The additional constraints ensure that parameters are consistent with the hierarchy assumptions of the model. The constraint sets $\mathcal{C}_3$ and $\mathcal{C}_4$ bound the same parameters as $\mathcal{C}_1$ and $\mathcal{C}_2$, but are more restrictive by replacing 0.3 by 0.2. Note that for these examples constraints were selected according to previous work [3]. Similarly to BKT$_C$, we ran the optimization ten times for each constraint set using different starting values. Subsequently, we selected the model with the highest likelihood on the training data set for comparison.

Prediction for the probabilistic models was performed as follows: we assumed the observation at time $t = 1$ to be given and predicted the outcome at time t with $t \in \{2, \ldots, T\}$ based on the previous $t - 1$ observations. To predict the outcome of the latent factors models, we evaluated the trained regression model at time t. The number of observations T for the different experiments is the minimal number of observations, which includes observations of all skills of the according model.

4.2 Data Sets and Models

We evaluated our DBN models on five large-scale data sets. To analyze the generalization capabilities of our approach, we selected data sets according to the following criteria: the data logs come from different ITS (`Calcularis`, `Andes2`, `Cognitive Tutor`, and `Dybuster`) and cover a wide range of age classes (from elementary school children over high school children up to university students) as well as different learning domains (mathematics, physics, spelling). In the following, we describe the five data sets as well as their respective DBNs. Note that we use the same skill sets for all our models, i.e. the skill sets described for the DBNs are also used for BKT$_C$, AFM, and PFM. We further assume that each observation is associated with exactly one skill.

Number Representation. To build the first model, we used data collected from `Calcularis`, an ITS for elementary school children with math learning difficulties [47]. `Calcularis` turns current neuro-cognitive theory into the design of different instructional games training number representations and number understanding as well as arithmetic operations. Student knowledge is represented as a DBN consisting of different mathematical skills [48], [49]. The data set contains log files of 1581 children with at least five sessions of 20 minutes per user.

The graphical model used in this experiment (see Fig. 1) is an excerpt of the skill model of `Calcularis` described in [48]. Skill S_A represents knowledge about the Arabic notation system. `Calcularis` does not contain any tasks associated with this skill. The ability to assign a number to an interval is denoted by S_G. The task associated with this skill is to guess a number in as few steps as possible. Finally, S_N denotes the ability to indicate the position of a number on a number line. We used a maximum of $T = 100$ observations per child for learning and prediction and specified the CPTs of the graphical model with $F = 22$ weights as illustrated in Fig. 1.

Subtraction. The second model is based on the same data set as the first model. This time, however, we investigated subtraction and number understanding skills. The graphical model (see Fig. 2) is again an excerpt of the skill model [48] of `Calcularis`. Subtraction skills are ordered according to their difficulty, which is determined by the magnitude of involved numbers, task complexity and the means allowed to solve a task. Skills S_1 (e.g., $48 - 6 = ?$), S_2 (e.g., $48 - 9 = ?$), S_3 (e.g., $48 - 26 = ?$), S_4 (e.g., $48 - 29 = ?$) and S_5 denote subtraction tasks in the number range $0 - 100$. We emphasize that there are no observation nodes associated with S_1 and S_5. The number understanding skill S_6 represents knowledge about the relational aspect of number (i.e., a number can be seen as a difference between two other numbers) in the number range $0 - 1{,}000$. Finally, skills S_7 (e.g., $158 - 3 = ?$), S_8 (e.g., $158 - 3 = ?$) and S_9 (e.g., $158 - 9 = ?$) represent subtraction tasks in the number range $0 - 1{,}000$. The difference between S_7 and S_8 lies in the means allowed to solve the task. A maximum of $T = 100$ observations per child was used for learning and prediction. Specification of the CPTs for the model requires $F = 86$ weights as illustrated in Fig. 2.

Physics. To build the third model, we used the 'USNA Physics Fall 2005' data set accessed via DataShop [50]. Data originate from 77 students of the United States Naval Academy and were collected from `Andes2`, an ITS for physics [18]. The tutor uses rule-based algorithms to build solution graphs that identify all possible solutions to the different tasks. Based on these graphs, a Bayesian network is constructed to assess the general physics knowledge of the student as well as the progress for the problem at hand. We used the different modules of the data set as skills for our experiment. The graphical model is depicted in Fig. 3. Note that, since we are neither experts regarding physics education nor concerning `Andes2`, we intentionally used a simplified model representing a subset of the modules as skills to avoid introducing incorrect assumptions. The model consists of the following modules: "Vectors" (S_V), "Translational Kinematics" (S_K), "Statistics" (S_S) and "Dynamics" (S_D). These modules consist of complex tasks for the given topic, i.e., calculating total forces in a system (see example in [18]). A maximum of $T = 500$ observations per child were considered for learning and prediction and the model can be described by $F = 33$ weights.

Algebra. This model is based on data from the KDD Cup 2010 Educational Data Mining Challenge (http://pslcdatashop.web.cmu.edu/KDDCup). The data set contains log files of 6,043 students that were collected by the `Cognitive Tutor` [4], an ITS for mathematics learning. The student model applied in this system is based on BKT, since the cognitive theory behind the cognitive tutor assumes skills to be independent.

We used the units of the 'Bridge to Algebra' course as skills for our experiment and selected four units of increasing difficulty, where students have to solve word problems

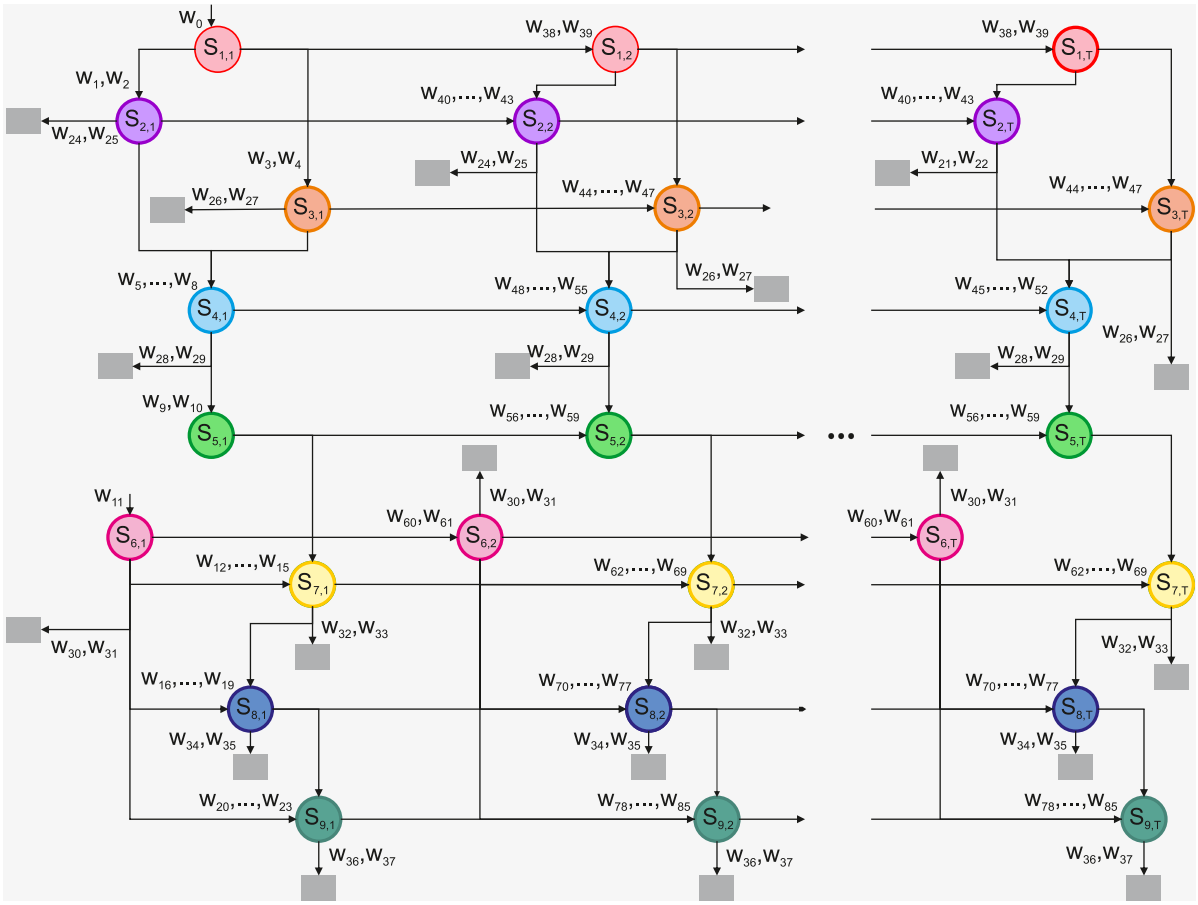

Fig. 2. Structure of the graphical model for the *Subtraction* data set. The DBN models the hierarchical relationships between the nine subtraction skills and contains 16 variables (nine latent and seven observable variables) per time step. The illustration shows the DBN unrolled over $T = 100$ observations.

involving calculations with whole numbers. The graphical model for this experiment is illustrated in Fig. 4. Skill S_A (e.g., $728{,}624 - 701{,}312$) denotes written addition and subtraction tasks without carrying/borrowing, while S_S involves carrying/borrowing (e.g., $728{,}624 - 703{,}303$). S_M (e.g., $33{,}564 \times 18$) and S_D (e.g., $10{,}810 \div 46$) represent long multiplications and divisions. The original model for the

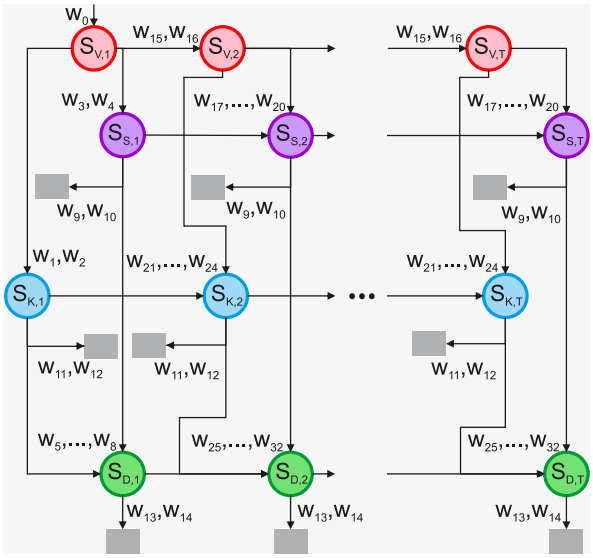

Fig. 3. Structure of the graphical model for the *Physics* data set. The DBN contains four latent nodes (the skills S_V, S_K, S_S, and S_D) and three observable variables per time step and evolves over $T = 500$ observations.

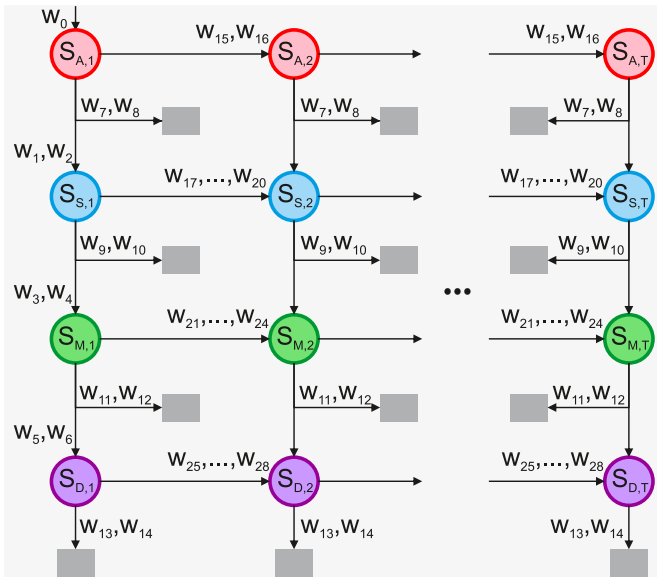

Fig. 4. Structure of the graphical model for the *Algebra* data set. The DBN represents the relationships between four skills (S_A, S_S, S_M, and S_D) and their associated observable nodes. The illustration shows the DBN unrolled over $T = 500$ observations.

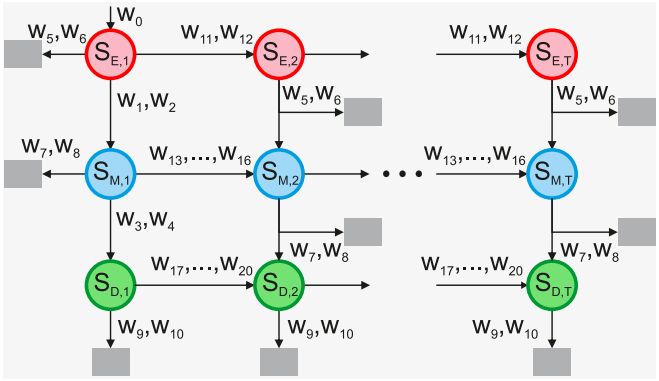

Fig. 5. Structure of the graphical model for the *Spelling* data set. The DBN consists of six variables per time step (three latent skill nodes and three observable nodes) and evolves over $T = 200$ observations.

data set used contains 94 skills that were carefully designed by domain experts. Building a fine-grained skill hierarchy would also require domain expert advice. Therefore, we again employed a simplified skill model representing a subset of the modules as skills. We used a maximum of $T = 500$ observations per student (consisting only of 'correct on first attempt responses') for learning and prediction and specified the CPTs of the model employing $F = 29$ weights.

Spelling. The last model uses data collected from `Dybuster`, an ITS for elementary school children with dyslexia [51]. The data set at hand contains data of 7,265 German-speaking children. `Dybuster` groups the words of a language into hierarchically ordered modules with respect to their frequency of occurrence in the language corpus as well as a word difficulty measure. The latter is computed based on the word length, the number of dyslexic pitfalls and the number of silent letters contained in the word.

We used these modules as skills to build our graphical model (see Fig. 5). Skills S_E, S_M and S_D denote the modules 3, 4 and 5 within `Dybuster`. Word examples are "warum" ("why", S_E), "Donnerstag" ("Thursday", S_M) and "Klapperschlange" ("rattlesnake", S_D). We used a maximum of $T = 200$ observations per child for the learning and prediction tasks and parametrized the model using $F = 21$ weights.

4.3 Prediction Accuracy

To answer our first research question *What are the benefits in terms of prediction accuracy of DBNs representing skill hierarchies compared to traditional student modeling approaches assuming independence between the different skills?*, we assessed the predictive performance of the different models using the root mean squared error (RMSE) and the area under the ROC curve (AUC). The RMSE is widely used in educational data mining to measure the performance of skill models and has demonstrated a high correlation to the 'moment of learning' (step in which a student learnt a skill) for BKT models [35], [52]. The AUC is also often applied for the evaluation of educational models [6], [27], [28], [29], [30], but its use as the only performance metric has been questioned lately [34]. However, in our case the AUC is used as an additional measure to the RMSE to assess the models' abilities to discriminate failures from successes. All error measures were calculated using 10-fold student-stratified cross validation. Statistical significance was computed using a two-sided t-test, correcting for multiple comparisons (Bonferroni-Holm). Table 1 details the RMSE and AUC for all models and data sets. Bold numbers denote a significant improvement of the constrained DBN model (for the different constraints $\mathcal{C}_1$ - $\mathcal{C}_4$ on the probabilities described in Section 4.1) over all alternative models (BKT$_C$, AFM and PFA). The best performing model for each measure is marked (*).

The constrained DBN approach yields significant and large improvements in prediction accuracy compared to BKT$_C$ for the *Number Representation* data set. We highlight the reduction of the RMSE by 3.8 percent and the large improvement achieved in AUC ($\text{AUC}_{BKT_C} = 0.5995$, $\text{AUC}_{\mathcal{C}_2} = 0.7026$). The constrained DBN models also outperform the logistic regression models in terms of RMSE and AUC with a reduction of the RMSE of 5.9 percent for AFM and 4.5 percent for PFA. AFM performance worst for both measures on this data set. While PFA exhibits an RMSE similar to BKT$_C$, it shows a far better AUC than BKT$_C$ ($\text{AUC}_{BKT_C} = 0.5995$, $\text{AUC}_{PFA} = 0.6717$). Note that all constrained DBN models significantly outperform all alternative models on both measures.

The resulting prediction accuracy for the *Subtraction* data set again demonstrates that the hierarchical DBN model outperforms the BKT$_C$ model as well as the logistic

TABLE 1
Prediction accuracy (RMSE and AUC) of the experiments, comparing different constraints sets for the DBNs with BKT$_C$, AFM, and PFA

		$DBN_{\mathcal{C}_1}$	$DBN_{\mathcal{C}_2}$	$DBN_{\mathcal{C}_3}$	$DBN_{\mathcal{C}_4}$	BKT$_C$	AFM	PFA
Number Representation	**RMSE**	**0.4469**	**0.4452**	**0.4416**	**0.4378***	0.4550	0.4657	0.4586
	AUC	**0.7008**	**0.7026***	**0.7010**	**0.6971**	0.5995	0.6262	0.6717
Subtraction	**RMSE**	0.4417	**0.4215***	0.4389	**0.4216**	0.4368	0.4457	0.4347
	AUC	0.6166	**0.6882**	0.6344	**0.6928***	0.5990	0.5763	0.6532
Physics	**RMSE**	0.4521	**0.4272**	**0.4244***	0.4465	0.4530	0.4527	0.4497
	AUC	**0.6644**	**0.7009**	**0.7021***	**0.6874**	0.4991	0.5425	0.5807
Algebra	**RMSE**	**0.3325**	**0.3246***	**0.3311**	**0.3260**	0.3369	0.3419	0.3374
	AUC	**0.6711**	**0.7042***	**0.6731**	**0.7040**	0.6012	0.6034	0.6407
Spelling	**RMSE**	0.4521	0.4495	0.4491	**0.4472***	0.4503	0.4498	0.4481
	AUC	0.5699	0.5775	0.5737	0.5808*	0.5029	0.5495	0.5790

Numbers in bold denote a significant improvement of the DBN model over all alternative models (BKT$_C$, AFM, and PFA). The best result for each performance measure is marked (*).

regression models. With a reduction of the RMSE by 3.5 percent compared to BKT_C and a reduction in RMSE of 3.0 percent over the PFA model, the benefits of the DBN model are again substantial. The DBN model exhibits a large improvement in AUC over BKT_C ($AUC_{BKT_C} = 0.5990$, $AUC_{\mathcal{C}_4} = 0.6928$). Note that the PFA model again significantly outperforms BKT_C regarding the AUC. The DBN model shows a large growth in AUC (compared to PFA) for the more restrictive constraint sets ($AUC_{\mathcal{C}_2} = 0.6882$, $AUC_{\mathcal{C}_4} = 0.6928$, $AUC_{PFA} = 0.6532$).

For the *Physics* data set, we observe significant and substantial improvements of the AUC over all alternative models ($AUC_{\mathcal{C}_3} = 0.7021$, $AUC_{BKT_C} = 0.4991$, $AUC_{PFA} = 0.5807$, $AUC_{AFM} = 0.5425$). Furthermore, the RMSE is reduced by 6.3 percent compared to BKT_C and by 6.6 percent compared to PFA. For this data set, the logistic regression models exhibit an RMSE similar to BKT_C, however, they show a substantially larger AUC than BKT_C.

On the *Algebra* data set, all DBN model configurations significantly outperform all other models regarding the RMSE and the AUC. We observe an improvement in RMSE of 3.7 percent compared to BKT_C and of 3.8 and 5.1 percent compared to PFA and AFM, respectively. The lower RMSE of all the models compared to the other data sets probably stems from the bias towards correct observations, i.e., this data set contains 85 percent of correct observations. Improvements in AUC are also substantial with all models exhibiting much lower values than DBN ($AUC_{\mathcal{C}_2} = 0.7042$, $AUC_{BKT_C} = 0.6012$, $AUC_{AFM} = 0.6034$, $AUC_{PFA} = 0.6407$).

Performance differences between the four models tend to be low for the *Spelling* data set. While still being significant, the absolute improvement in RMSE of the DBN models compared to BKT_C drops to 0.1 percent. The logistic regression models show a similar performance ($RMSE_{PFA} = 0.4481$, $RMSE_{AFM} = 0.4498$). The DBN models again outperform AFM and BKT_C regarding the AUC ($AUC_{\mathcal{C}_4} = 0.5808$, $AUC_{BKT_C} = 0.5029$, $AUC_{PFA} = 0.5495$) and perform in range with PFA ($AUC_{PFA} = 0.5790$). However, AUC values for all the models tend to be lower than the values achieved on the other data sets.

4.4 Model Behavior

Our second research question *Where do the differences in prediction accuracy between the different student models come from?* aims at understanding model behavior in more detail and investigating the observed differences in prediction accuracy (see Section 4.3) between the different approaches. To answer this question, we computed the Brier score of the different models [34]. The Brier score is equivalent to the mean squared error (MSE) and can be decomposed into three components: BS = REL − RES + UNC. The uncertainty (UNC) quantifies the inherent uncertainty of the events (correct and incorrect outcomes) in the data set and is independent of the selected model. The reliability term (REL) measures the difference between predicted and observed probabilities. A reliability of zero ($REL = 0$) indicates a perfectly reliable prediction. The resolution term (RES) measures how much predictions differ from the base rate (proportion of correct outcomes in the data set) and therefore gives and indication of the range of predicted probabilities, i.e., a high resolution is desirable. An ideal model would therefore minimize the REL term, while maximizing the RES term. We computed the Brier score and its components using 10-fold student-stratified cross validation. Statistical significance was calculated using a two-sided t-test, correcting for multiple comparisons (Bonferroni-Holm). The reliability and resolution for the different data sets are displayed in Fig. 6. For the DBNs, we included only the model with the best RMSE for each data set (see Table 1).

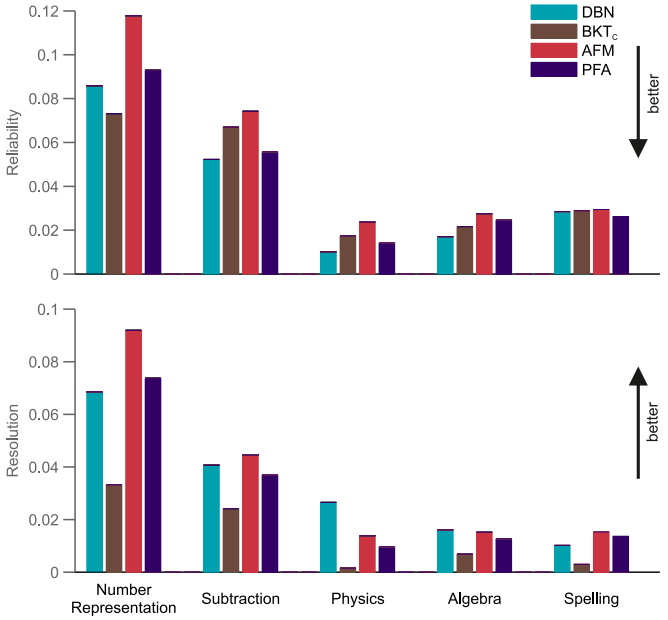

Fig. 6. Reliability (REL) and Resolution (RES) over the different data sets for BKT_C, AFM, PFA, and the DBN model with the lowest RMSE.

For the *Number Representation* data set, BKT_C exhibits a significantly better reliability than the DBN model with the best RMSE ($REL_{BKT_C} = 0.0732$, $REL_{\mathcal{C}_4} = 0.0859$). The reliabilities of the DBN model and the PFA model are in range ($REL_{PFA} = 0.0929$), while AFM shows a much worse reliability ($REL_{AFM} = 0.1178$), but has the best resolution ($RES_{AFM} = 0.0919$). Resolutions of the DBN and PFA models are again similar ($RES_{PFA} = 0.0737$, $RES_{\mathcal{C}_4} = 0.0685$). For this data set, BKT_C achieves a significantly lower resolution than DBN ($RES_{BKT_C} = 0.0332$), which explains the performance differences in RMSE.

We observe a different picture for the *Subtraction* data set. While the DBN and PFA models show again similar reliabilities ($REL_{\mathcal{C}_2} = 0.0522$, $REL_{PFA} = 0.0557$), BKT_C exhibits a worse reliability ($REL_{BKT_C} = 0.0671$); however, the differences in reliability between BKT_C and DBN are not statistically significant. The DBN model again demonstrates a significantly higher resolution than BKT_C ($RES_{\mathcal{C}_2} = 0.0406$, $RES_{BKT_C} = 0.0241$). Also the results for the AFM model are confirmed, which shows the worst reliability ($REL_{AFM} = 0.0741$), but a high resolution ($RES_{AFM} = 0.0445$).

On the *Physics* data set, the BKT_C, PFA, and DBN models shows similar reliabilities ($REL_{\mathcal{C}_3} = 0.0127$, $REL_{PFA} = 0.0141$, $REL_{BKT_C} = 0.0174$), the AFM again demonstrates a significantly worse reliability than the DBN model ($REL_{AFM} = 0.0237$). For this data set, the DBN model exhibits a significantly higher resolution than all the other models ($REL_{\mathcal{C}_3} = 0.0262$, $RES_{BKT_C} = 0.0016$, $RES_{PFA} = 0.0096$, $RES_{AFM} = 0.0138$).

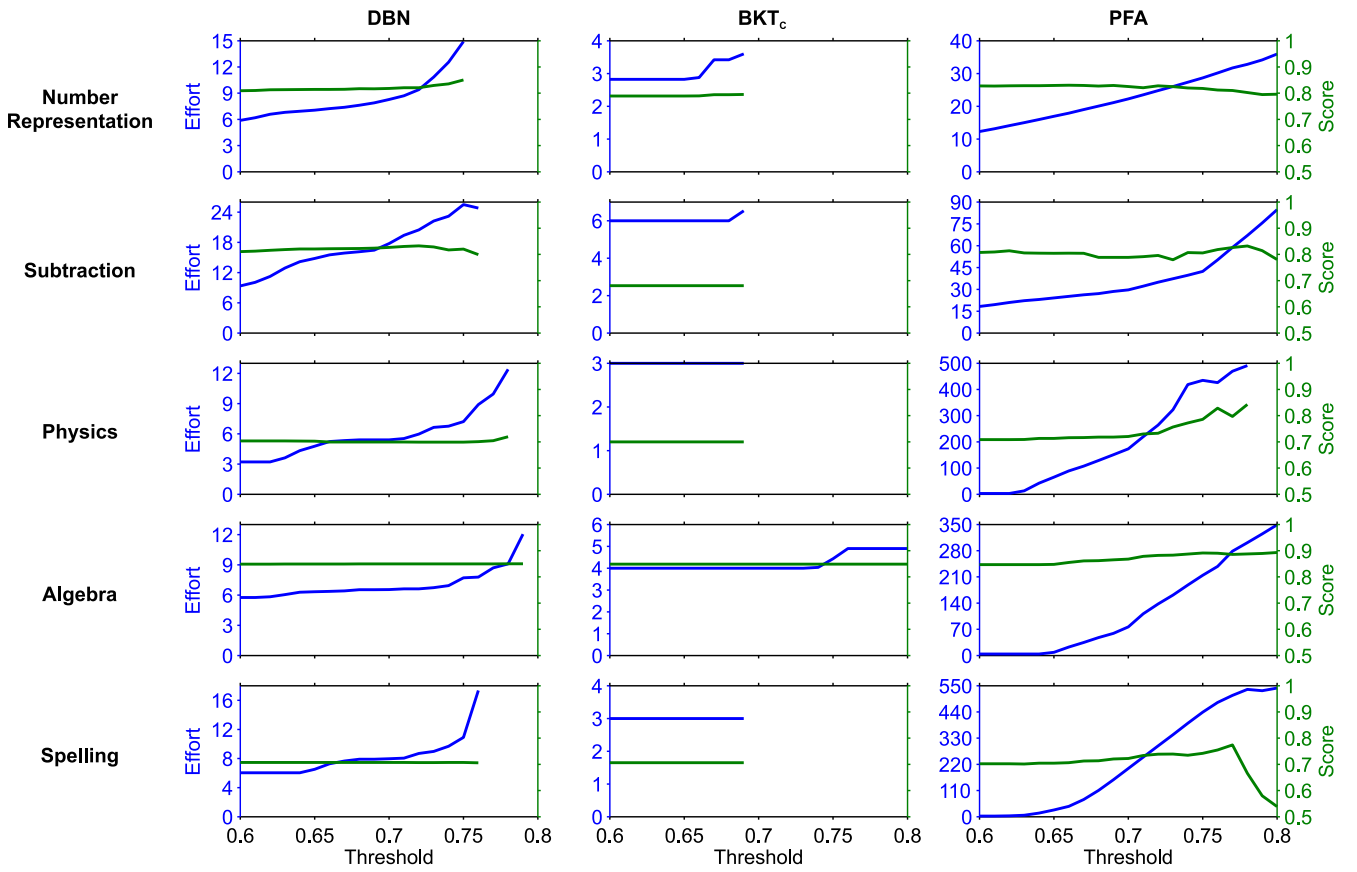

Fig. 7. Effort (blue) and score (green) for the DBN model with the lowest RMSE (left), BKT$_C$ (middle), and PFA (right) over all different data sets. The x-axes denote the thresholds R, the y-axes denote the effort (left *y*-axis) and score (right *y*-axis), respectively.

The DBN model achieves a better reliability than all the other models on the *Algebra* data set, with all differences being statistically significant (REL$_{\mathcal{C}_2} = 0.0170$, REL$_{PFA} = 0.0246$, REL$_{BKT_C} = 0.0216$, REL$_{AFM} = 0.0275$). The same fact holds true for the resolution, where the DBN model again significantly outperforms all the other models (RES$_{\mathcal{C}_2} = 0.0160$, RES$_{BKT_C} = 0.0069$, RES$_{PFA} = 0.0127$, RES$_{AFM} = 0.0152$). This result explains the lower RMSE achieved for all tested DBN models (see Table 1).

All models achieve similar reliability on the *Spelling* data set (REL$_{\mathcal{C}_4} = 0.0284$, REL$_{PFA} = 0.0262$, REL$_{BKT_C} = 0.0288$, REL$_{AFM} = 0.0294$), with all the small differences between DBN and the other models being statistically significant. There is no significant difference between the resolution of the DBN and the PFA models (RES$_{\mathcal{C}_4} = 0.0102$, RES$_{PFA} = 0.0137$). The AFM achieves a significantly higher resolution than the DBN model (RES$_{AFM} = 0.0153$), while the DBN model significantly outperforms BKT$_C$ regarding the resolution (RES$_{BKT_C} = 0.0030$).

4.5 Instructional Policies

Lately, the influence of small differences in performance measures on the learning outcome has been questioned [25]. However, it has been shown that models with similar predictive performance can lead to significantly different instructional policies [1], [36], which leads to our third research question: *How do the different student modeling techniques influence instructional policies?*. To evaluate the impact of the different models on the task selection, we calculated the effort E (number of practice opportunities needed to pass a skill) and score S (percentage of correct observations after having the skill passed) of the different models. We used the *Teal* framework [36] to define the mastery criterion: Instruction for a skill s is stopped, when the predicted probability $\hat{p}_t$ of correctly answering the next task t associated with skill s is above a given threshold R, i.e., $\hat{p}_t > R$. Since the effort and score are dependent on the threshold R employed, we computed the measures for a range of thresholds $R \in [0.6, \ldots, 0.8]$. We again calculated all measures using 10-fold student-stratified cross validation. We included only one logistic regression model (PFA) in this evaluation, since PFA consistently outperformed AFM on the performance measures over all data sets. Furthermore, we again only included the DBN model with the best RMSE (see Table 1). Fig. 7 shows the resulting effort (blue) and score (green) for DBN, BKT$_C$ and PFA at different thresholds R over all data sets.

On the *Number Representation* data set, the three models show similar scores. The maximum score ($S = 0.85$) for the DBN model is reached at $R = 0.75$ with an average effort $E = 14.92$. An effort of $E = 14.92$ can be considered as large, since it meets the wheel-spinning criterion (effort $E > 10$ [53]). However, the model still reaches a score $S = 0.82$ with an average effort below this limit ($E = 9.4$). The BKT$_C$ model achieves the lowest maximum score $S = 0.79$ of all the models, however, the average effort $E = 3.60$ is also low. Note that effort and score for the BKT$_C$ model could be computed only for $R \leq 0.69$ since the model never predicted a probability $\hat{p}_t > 0.69$. The values for the scores show very little differences, this behavior can be

explained by the low resolution of the BKT_C model (see Fig. 6, bottom) on this data set. The PFA model exhibits the maximum score ($S = 0.83$) at $R = 0.66$ with the highest average effort of $E = 17.90$. The PFA model generally predicts flatter learning curves on this data set, resulting in the higher effort E needed for the given thresholds.

For the *Subtraction* data set, the results look similar. The BKT_C model exhibits a flat score curve (maximum score $S = 0.68$), but also the lowest average effort ($E = 6.00$). The DBN model again achieves a considerably higher score ($S = 0.81$) than the BKT_C model with a reasonable effort $E = 10.07$. The PFA model predicts student performance much more conservatively than the other models, resulting in high efforts necessary to reach $\hat{p}_t > R$: The PFA model achieves scores around $S = 0.8$ for an average effort $E > 15$.

On the *Physics* data set, effort ($E = 3.00$) and score ($S = 0.70$) take just one value for the BKT_C model. This is due to the small resolution ($RES_{BKT_C} = 0.0016$) of the BKT_C model on this data set (illustrated in Fig. 6, bottom). For the PFA model, scores increase with a higher effort, but the model generally exhibits very high efforts. However, for an effort $E = 3.00$, the model achieves a score $S = 0.71$, which is in the range of the BKT_C model. Also the DBN model achieves a score $S = 0.70$ with an effort $E = 3.21$ on average. Similarly to the BKT_C model, scores do not increase with higher effort.

For the *Algebra* data set, all models achieve higher scores ($S_{BKT_C} = 0.85$, $S_{PFA} = 0.85$, $S_{C_2} = 0.85$) than on the other data sets with similar efforts ($E_{BKT_C} = 4.00$, $E_{PFA} = 4.00$, $E_{C_2} = 5.75$). This observation is probably due to the fact that the *Algebra* data set contains easier skills; this bias towards correct observations is also visible in the low RMSE of all models for this data set (see Table 1). For the PFA model, scores get as high as $S = 0.90$, but also the average effort grows to $E = 349.27$. This high number of observations is possible due to the simplified skill model: we used the units of the course as skills to build the model.

The effort and score for the different models on the *Spelling* data set confirm the previous results: We observe similar scores ($S_{BKT_C} = 0.71$, $S_{PFA} = 0.70$, $S_{C_4} = 0.71$) with reasonable efforts ($E_{BKT_C} = 3.00$, $E_{PFA} = 3.00$, $E_{C_4} = 6.00$) for the three models. The BKT_C model again takes only one value for effort and score. For the DBN model, the scores do not increase with ascending effort. The scores for the PFA model slowly increase over time, however, they drop for very high efforts.

5 Discussion

The student model is an integral part of an ITS. One of the most popular approaches to student modeling is BKT. However, this technique cannot model the relationships between the different skills of a learning domain. More complex DBN models have the potential to increase the representational power by modeling skill hierarchies. In this work, we have suggested the use of DBNs representing skill hierarchies for student modeling and have extensively evaluated them on five large-scale data sets. Our experiments answer three important research questions.

The first research question analyzed the prediction accuracy of our approach: *What are the benefits in terms of prediction accuracy of DBNs representing skill hierarchies compared to traditional student modeling approaches assuming independence between the different skills?*. Our results demonstrate that DBN models outperform BKT in prediction accuracy. For hierarchical domains, the RMSE can be reduced by 3.5 percent (*Subtraction* data set) to 6.3 percent (*Physics* data set). As expected, adding skill topologies has a much smaller benefit for learning domains that are less hierarchical in nature (such as spelling learning). While differences in RMSE are still statistically significant, the magnitude of improvement is low. The results obtained on the *Physics* and *Algebra* data sets show that even simple hierarchical models improve prediction accuracy significantly. A domain expert employing a more detailed skill topology and more complex constraint sets could probably obtain an even higher accuracy on these data sets. Our comparison to logistic regression models also shows the superiority of the DBN models: The DBN models significantly outperform PFA on all data sets. PFA exhibits a similar or better RMSE than BKT_C on the data sets. These results are not unexpected since PFA has been shown to be superior to AFM and BKT [2]. However, these findings have been disputed for BKT previously [29], [54]. Note that all models exhibit a much lower RMSE on the *Algebra* data set than on the other data sets. This better performance results from the bias in the data set towards correct observations (85 percent of correct observations): prediction accuracy is related to the bias of the data set, i.e., a uniform distribution of correct and wrong observations is most difficult to predict. The significant reduction obtained in RMSE is very promising. The RMSE is a proper score [55] and has been widely used in educational data mining to measure student model performance, for example [7], [10], [11], [25]. The model fitting procedures we used optimize the log-likelihood of the data; the RMSE shows a high correlation to this measure [34]. Furthermore, the RMSE is also correlated to the 'moment of learning' (step in which a student learnt a skill) [35], [52].

We also computed the AUC of the different models. While the AUC was criticized as a performance measure lately [34], [35], it is a useful additional measure (for example in combination with the RMSE) to assess the models' ability to discriminate failure from success. In a biased data set with a high base rate (i.e., a high frequency of correct outcomes), a model always outputting the base rate performs well regarding the RMSE, since the prediction of the base rate achieves perfect reliability. However, such a model exhibits a low AUC. The DBN models consistently show the highest AUC over all hierarchical data sets. For the *Spelling* data sets, DBN and PFA perform similarly. BKT_C shows a significantly lower AUC than the DBN models.

For a better understanding of model behavior, a more detailed quantification of model performance is useful. To answer our second research question *Where do the differences in prediction accuracy between the different student models come from?*, we therefore computed the Brier score of the models [34]. The BKT_C model consistently showed the lowest resolution (RES) of all models and was outperformed by the DBN models over all data sets. This result is not

unexpected: The AUC is related to the Brier score [56] and the BKT models exhibited the lowest AUC across all data sets of our experiments. Furthermore, these results are in line with previous work which found that BKT models tend to have a low resolution (RES) [34]. We observe mixed results for the reliability: BKT_C exhibited a significantly better reliability than the DBN model on the *Number Representation* data set and a significantly worse reliability than DBN on the *Algebra* and *Spelling* data sets—reliabilities were similar on all other data sets. We therefore conclude that the differences in RMSE between BKT_C and DBN models are mainly due to the low resolution of BKT_C. The DBN models performed similarly or better in reliability and resolution than PFA (with the exception of the *Spelling* data set) and therefore the differences in RMSE between these two models probably arise from the combination of the two measures. The AFM consistently showed a bad reliability (worst model over all data sets), but performed very well regarding the resolution. The bad reliability of the AFM might stem from the fact, that AFM does not differentiate between correct and wrong observations. The superior resolution of the logistic regression models is probably due to their nature, i.e., they are fitting a learning curve over time. Our analyses demonstrate that a detailed quantification of model performance is helpful for understanding the properties of the different models.

Lately, the evaluation of student models based on small differences in performance metrics has been questioned [25]. However, previous work has demonstrated, that small differences in RMSE can have a large impact on over-practice [7], [40]. Furthermore, the improvement in performance metric has a significant impact on the expected number of practice opportunities needed [57] and the instructional policy of the system [1]. Other research discussed the use of a single number as a performance metric [34], [36], [37]. We therefore used the *Teal* framework [36] to answer our third research question: *How do the different student modeling techniques influence instructional policies?*. *Teal* is threshold dependent and can be applied to all models used in our experiments.

The BKT_C models consistently showed a low effort over all data sets, scores were in range with the other models. Notice, however, that BKT_C demonstrated a very low variance in effort and scores both over time and over the different data sets. This low variance is caused by the low resolution (and AUC) of BKT_C. While the generally low efforts and scores might seem promising, the low variance limits instructional design, since the effort and score cannot be tuned by adapting the threshold. The DBN models tended to be more pessimistic than BKT_C, i.e., they generally required a higher effort to reach the same score. This is due to the fact that these models represent forgetting ($p_f \geq 0$), which leads to a slower increase of the predicted probability $\hat{p}_t$. Nevertheless, the DBN models reached a similar or higher score than BKT_C with a reasonable effort $E < 10$ (students with an effort $E > 10$ can be considered to be wheel-spinning) [53]. The PFA models showed similar effort and scores as BKT_C for lower thresholds, but effort increased very fast with not much gain regarding the score. These results are in line with the observations made in previous work [1]. We conclude from our experiments that DBN models yield meaningful instructional policies relating to cognitive mastery.

From Fig. 7, it is very well visible, that *Teal* is threshold dependent. Different thresholds are optimal for the different models. This dependency on a threshold gives flexibility to instructional design, since the resulting policies can be tuned using the threshold. However, the optimal choice of threshold might be difficult, but could be done in a data-driven way by first determining a target score. Therefore, methods not requiring a threshold are an interesting alternative. The predictive similarity policy [1] for example stops when the predicted probabilities do not change anymore regardless of the student's answer (correct or wrong).

The results of our extensive experimental evaluation demonstrate that DBNs are a valuable technique for student modeling. Our DBN models show a better RMSE and AUC than traditional student models such as PFA and BKT over a wide range of data sets. Furthermore, the DBN models yield meaningful instructional policies. Note, however, that the performance differences between DBN and BKT, especially the influence of the different parameters, need to be investigated further. Recent work [58] has demonstrated that the prediction accuracy of BKT can be significantly improved when allowing forgetting ($p_F > 0$). Furthermore, the reported results on the real world data sets used are only approximations to the true prediction accuracy and performance of the instructional policies. The data sets stem from learning systems with student models and instructional policies already in place, resulting in the introduction of a bias into the generated log data.

6 Conclusion

In this work, we showed that DBN models are well suited for representing student knowledge. We extensively evaluated our approach on five large data sets from different learning domains containing students of a wide age range. The results demonstrate that the incorporation of skill topologies yields significant improvement in prediction accuracy over traditional student models. Furthermore, the more complex hierarchical DBN models also lead to meaningful and interpretable instructional policies. The use of the same parametrization and constraint sets for all experiments demonstrates that basic assumptions about learning hold across different learning domains. To conclude, our results show that modeling skill topologies is beneficial and easy to use, as even simple hierarchies and parametrization lead to significant improvements in prediction accuracy and instructional policies.

In the future, we would like to analyze the influence of the skill hierarchies, the different parameters and the reasons for the performance differences between the different approaches further. We furthermore plan to apply the individualization techniques used in BKT [7], [10], [11] to DBNs and compare the benefits of introducing skill hierarchies to the advantages of student-individualized modeling techniques. Moreover, we would like to explore the possibility of learning skill hierarchies from data.

Acknowledgments

This work was supported by ETH Research Grant ETH-23 13-2.

REFERENCES

[1] J. Rollinson and E. Brunskill, "From predictive models to instructional policies," in *Proc. Int. Conf. Educational Data Mining*, 2015.

[2] P. I. Pavlik, H. Cen, and K. R. Koedinger, "Performance factors analysis - A new alternative to knowledge tracing," in *Proc. Conf. Artif. Intell. Educ.: Building Learn. Syst. Care: Knowl. Representation Affective Modelling*, 2009, pp. 531–538 .

[3] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Model. User-Adapted Interaction*, vol. 4, pp. 253–278, 1994.

[4] K. R. Koedinger, J. R. Anderson, W. H. Hadley, and M. A. Mark, "Intelligent tutoring goes to school in the big city," *Int. J. Artif. Intell. Educ.*, 1997, pp. 30–43.

[5] J. E. Beck and J. Sison, "Using knowledge tracing to measure student reading proficiencies," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2004, pp 624–634.

[6] R. S. Baker, et al., "Contextual slip and prediction of student performance after use of an intelligent tutor," in *Proc. Int. Conf. User Model. Adaptation Personalization*, 2010, pp. 52–63.

[7] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon, "Individualized bayesian knowledge tracing models," in *Proc. 16th Int. Conf. Artif. Intell. Educ.*, 2013.

[8] K.-M. Chang, J. Beck, J. Mostow, and A. Corbett, "A bayes net toolkit for student modeling in intelligent tutoring systems," in *Proc. Bayes Net Toolkit Student Model. Intell. Tutoring Syst.*, 2006, pp. 104–113.

[9] Z. A. Pardos, S. Trivedi, N. T. Heffernan, and G. N. Sárközy, "Clustered knowledge tracing," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2012, pp. 405–410.

[10] Z. A. Pardos and N. T. Heffernan, "Modeling individualization in a bayesian networks implementation of knowledge tracing," in *Proc. 18th Int. Conf. User Model. Adaptation Personalization*, 2010, 255–266.

[11] Y. Wang and N. T. Heffernan, "The student skill model," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2012, pp. 399–404.

[12] Y. Wang and J. Beck, "Class versus student in a Bayesian network student model," in *Proc. Int. Conf. Artif. Intell. Educ.*, 2013, pp. 151–160.

[13] K. T. Wang, Y.-M. Huang, Y.-L. Jeng, and T.-I. Wang, "A blog-based dynamic learning map," *Comput. Educ.*, vol. 51, pp. 262–278, 2008.

[14] J. P. Leighton, M. J. Gierl, and S. M. Hunka, "The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach," *J. Educational Meas.*, vol. 41, pp. 205–237, 2004.

[15] E. Millán, T. Loboda, and J. L. P. de-la Cruz, "Bayesian networks for student model engineering," *Comput. Educ.*, vol. 55, pp. 1663–1683, 2010.

[16] E. Millán, L. Descalço, G. Castillo, P. Oliveira, and S. Diogo, "Using Bayesian networks to improve knowledge assessment," *Comput. Educ.*, vol. 60, pp. 436–447, 2013.

[17] P. García, A. Amandi, S. Schiaffino, and M. Campo, "Evaluating Bayesian networks precision for detecting students learning styles," *Comput. Educ.*, vol. 49, pp. 794–808, 2007.

[18] C. Conati, A. Gertner, and K. VanLehn, "Using Bayesian networks to manage uncertainty in student modeling," *User Model. User-Adapted Interaction*, vol. 12, pp. 371–417, 2002.

[19] M. Mayo and A. Mitrovic, "Optimising ITS behaviour with Bayesian networks and decision theory," *Int. J. Artif. Intell. Educ.*, vol. 12, pp. 124–153, 2001.

[20] G.-M. Baschera, A. G. Busetto, S. Klingler, J. Buhmann, and M. Gross, "Modeling engagement dynamics in spelling learning," in *Proc. Int. Conf. Artif. Intell. Educ.*, 2011, pp. 31–38.

[21] T. Käser, et al., "Towards a framework for modelling engagement dynamics in multiple learning domains," *Int. J. Artif. Intell. Educ.*, vol. 22, pp. 42–70, 2012.

[22] E. Horvitz, J. Breese, D. Heckerman, D. Hovel, and K. Rommelse, "The lumière project: Bayesian user modeling for inferring the goals and needs of software users," in *Proc. 14th Conf. Uncertainty Artif. Intell.*, 1998, pp. 256–265.

[23] M. C. Frank and J. B. Tenenbaum, "Three ideal observer models for rule learning in simple languages," *Cognition*, vol. 120, pp. 360–71, 2010.

[24] C. Baker, J. B. Tenenbaum, and R. Saxe, "Bayesian models of human action understanding," in *Proc. Advances Neural Inf. Process. Syst.*, 2005.

[25] J. Beck and X. Xiong, "Limits to accuracy: How well can we do at student modeling?" in *Proc. 6th Int. Conf. Educational Data Mining*, 2013.

[26] Y. Gong, J. E. Beck, and N. T. Heffernan, "Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2010, pp. 35–44.

[27] R. S. Baker, A. T. Corbett, and V. Aleven, "More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2008, pp. 406–415.

[28] J. E. Beck and K. M. Chang, "Identifiability: A fundamental problem of student modeling," in *Proc. Int. Conf. User Model.*, 2007, pp. 137–146.

[29] Z. A. Pardos, S. M. Gowda, R. S. Baker, and N. T. Heffernan, "The sum is greater than the parts: Ensembling models of student knowledge in educational software," *SIGKDD Explorations Newsletter*, vol. 13, pp. 37–44, 2012.

[30] J. P. González-Brenes, Y. Huang, and P. Brusilovsky, "General features in knowledge tracing: Applications to multiple subskills, temporal item response theory, and expert knowledge," in *Proc. 7th Int. Conf. Educational Data Mining*, 2014.

[31] H. Cen, K. Koedinger, and B. Junker, "Learning factors analysis – A general method for cognitive model evaluation and improvement," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2006, pp. 164–175.

[32] Z. A. Pardos and N. T. Heffernan, "Modeling individualization in a Bayesian networks implementation of knowledge tracing," in *Proc. 18th Int. Conf. User Model. Adaptation Personalization.*, 2010, pp. 255–266.

[33] M. Khajah, R. M. Wing, R. V. Lindsey, and M. C. Mozer, "Integrating latent-factor and knowledge-tracing models to predict individual differences in learning," in *Proc. Int. Conf. Educational Data Mining*, 2014.

[34] R. Pelánek, "Metrics for evaluation of student models," *J. Educational Data Mining*, vol. 7, 2015.

[35] Z. A. Pardos and M. V. Yudelson, "Towards moment of learning accuracy," in *Proc. Int. Conf. Artif. Intell. Educ. Workshop*, 2013.

[36] J. P. González-Brenes and Y. Huang, "Your model is predictive but is it useful? Theoretical and empirical considerations of a new paradigm for adaptive tutoring evaluation," in *Proc. 8th Int. Conf. Educational Data Mining*, 2015.

[37] Y. Huang, J. P. González-Brenes, R. Kumar, and P. Brusilovsky, "A framework for multifaceted evaluation of student models," in *Proc. Int. Conf. Educational Data Mining Soc.*, 2015, Art. no. ED560517.

[38] T. Käser, A. G. Schwing, T. Hazan, and M. Gross, "Computational education using latent structured prediction," *Proc. 17th Int. Conf. Artif. Intell. Statist.*, 2014, pp. 540–548.

[39] T. Käser, S. Klingler, A. G. Schwing, and M. Gross, "Beyond Knowledge Tracing: Modeling Skill Topologies with Bayesian Networks," in *Proc. Intell. Tutoring Syst.*, 2014, pp. 188–198.

[40] H. Cen, K. R. Koedinger, and B. Junker, "Is over practice necessary? -Improving learning efficiency with the cognitive tutor through educational data mining," in *Proc. Conf. Artif. Intell. Educ.*, 2007, pp. 511–518.

[41] H. Cen, K. R. Koedinger, and B. Junker, "Comparing two IRT models for conjunctive skills," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2008, pp. 796–798.

[42] J. Reye, "Student modelling based on belief networks," *Int. J. Artif. Intell. Educ.*, vol. 14, pp. 63–96, 2004.

[43] N. Friedman, K. Murphy, and S. Russell, "Learning the structure of dynamic probabilistic networks," in *Proc. 14th Conf. Uncertainty Artif. Intell.*, 1998, pp. 139–147.

[44] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 282–289.

[45] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun, "Efficient structured prediction with latent variables for general graphical models," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012.

[46] Z. A. Pardos and N. T. Heffernan, "Navigating the parameter space of Bayesian knowledge tracing models: Visualizations of the convergence of the expectation maximization algorithm," in *Proc. 3rd Int. Conf. Educational Data Mining*, 2010.

[47] T. Käser, et al., "Design and evaluation of the computer-based training program calcularis for enhancing numerical cognition," *Front. Psychol.*, vol. 4, 2013, Art. no. 489.

[48] T. Käser, et al., "Modelling and optimizing mathematics learning in children," *Int. J. Artif. Intell. Educ.*, vol. 23, pp. 115–135, 2013.

[49] T. Käser, et al., "Modelling and optimizing the process of learning mathematics," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2012, pp. 389–398.

[50] K. Koedinger, et al., "A Data Repository for the EDM community: The PSLC DataShop," in *Handbook of Educational Data Mining*. Boca Raton, FL, USA: CRC Press, 2010.
[51] M. Gross and C. Vögeli, "A multimedia framework for effective language training," *Comput. Graph.*, vol. 31, pp. 761–777, 2007.
[52] R. S. J. D. Baker, A. B. Goldstein, and N. T. Heffernan, "Detecting learning moment-by-moment," *Int. J. Artificial Intell. Educ.*, vol. 21, pp. 5–25, 2011.
[53] J. E. Beck and Y. Gong, "Wheel-spinning: Students who fail to master a skill," in *Proc. Int. Conf. Artif. Intell. Educ.*, 2013, pp. 431–440.
[54] R. S. J. D. Baker, Z. A. Pardos, S. M. Gowda, B. B. Nooraei, and N. T. Heffernan, "Ensembling predictions of student knowledge within intelligent tutoring systems," *Proc. Int. Conf. User Model. Adaptation Personalization*, 2011, pp. 13–24.
[55] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. American Statist. Assoc.*, vol. 102, 2007, Art. no. 477.
[56] J. Hernández-Orallo, P. Flach, and C. Ferri, "A unified view of performance metrics: Translating threshold choice into expected classification loss," *J. Mach. Learn. Res.*, vol. 13, pp. 2813–2869, 2012.
[57] J. I. Lee and E. Brunskill, "The impact on individualizing student models on necessary practice opportunities," in *Proc. 5th Int. Conf. Educational Data Mining*, 2012, pp. 118–125.
[58] M. Khajah, R. V. Lindsey, and M. C. Mozer, "How deep is knowledge tracing?" in *Proc. Int. Conf. Educational Data Mining*, 2016.