On the Performance Characteristics of Latent-Factor and Knowledge Tracing Models

Severin Klingler
Department of Computer Science
ETH Zurich, Switzerland
kseverin@inf.ethz.ch

Tanja Käser
Department of Computer Science
ETH Zurich, Switzerland
kaesert@inf.ethz.ch

Barbara Solenthaler
Department of Computer Science
ETH Zurich, Switzerland
sobarbar@inf.ethz.ch

Markus Gross
Department of Computer Science
ETH Zurich, Switzerland
grossm@inf.ethz.ch

ABSTRACT

Modeling student knowledge is a fundamental task of an intelligent tutoring system. A popular approach for modeling the acquisition of knowledge is Bayesian Knowledge Tracing (BKT). Various extensions to the original BKT model have been proposed, among them two novel models that unify BKT and Item Response Theory (IRT). Latent Factor Knowledge Tracing (LFKT) and Feature Aware Student knowledge Tracing (FAST) exhibit state of the art prediction accuracy. However, only few studies have analyzed the characteristics of these different models. In this paper, we therefore evaluate and compare properties of the models using synthetic data sets. We sample from a combined student model that encompasses all four models. Based on the true parameters of the data generating process, we assess model performance characteristics for over 66'000 parameter configurations and identify best and worst case performance. Using regression we analyze the influence of different sampling parameters on the performance of the models and study their robustness under different model assumption violations.

Keywords

Knowledge Tracing, Item Response Theory, synthetic data, predictive performance, robustness

1. INTRODUCTION

A fundamental part of an intelligent tutoring system (ITS) is the student model. Task selection and evaluation of the student's learning progress are based on this model, and therefore it influences the learning experience and the learning outcome of a student. Thus, accurately modeling and predicting student knowledge is essential.

Approaches for student modeling are usually based on two popular techniques: Item Response Theory (IRT) [36] and Bayesian Knowledge Tracing (BKT) [9]. The concept of IRT assumes that that the probability of a correct response to an item is a mathematical function of student and item parameters. The Additive Factors Model (AFM) [7, 8] fits a learning curve to the data by applying a logistic regression. Another technique called Performance Factors Analysis (PFA) [27] is based on the Rasch item response model [12]. BKT models student knowledge as a binary variable that can be inferred by binary observations. Performance of the original BKT model has been improved by using individualization techniques such as modeling the parameters by student and skill [23, 35, 39] or per school class [34]. Clustering approaches [25] have also proven successful in improving the prediction accuracy of BKT. Furthermore, hybrid models combining the approaches of IRT and BKT have been proposed. In [17] a dynamic mixture model has been presented to trace performance and affect simultaneously. The KT-IDEM model extends BKT by introducing item difficulty parameters [22]. Other work focused on individualizing the initial mastery probability of BKT by using IRT [38]. Logistic regression has also been used to integrate subskills into BKT [37]. Recently, two models have been introduced which synthesize IRT and BKT. Latent Factor Knowledge Tracing (LFKT) [18] individualizes the guess and slip probabilities of BKT based on student ability and item difficulty. Feature Aware Student Knowledge Tracing (FAST) [14] generalizes the individualized guess and slip probabilities to arbitrary features.

Lately, the analysis of properties of BKT has gained increasing attention. It has been shown [5] that learning BKT models exhibits fundamental identifiability problems, i.e., different model parameter estimates may lead to identical predictions about student performance. This problem was addressed by using an approach that biases the model search by Dirichlet priors to get statistically reliable improvements in predictive performance. [33] extended this work by performing a fixed point analysis of the solutions of the BKT learning task and by deriving constraints on the range of parameters that lead to unique solutions. Furthermore, it has been shown that the parameter space of BKT models

can be reduced using clustering [30]. Other research focused on analyzing convergence properties [24] of the expectation maximization algorithm (EM) for learning BKT models and exploring parameter estimates produced by EM [15]. It has been shown that convergence in the log likelihood space does not necessarily mean convergence in the parameter space. [11] have studied how good BKT is at predicting the moment of mastery. Different thresholds to assess mastery and their corresponding lag, i.e., the number of tasks that BKT needs to assess mastery (after mastery has already been achieved), have been investigated. Using multiple model fitting procedures, BKT has been compared to PFA [13]. While no differences in predictive accuracy between the models have been reported, it has been shown that for knowledge tracing EM achieves significantly higher predictive accuracy than Brute Force. Findings from other studies, however, suggest the opposite [1, 2]. In [4], upper bounds on the predictive performance have been investigated by employing various cheating models. It has been concluded that BKT and PFA perform close to these limits, suggesting that other factors such as robust learning or optimal waiting intervals should be considered to improve tutorial decision making. The predictive performance of LFKT and FAST has been compared to KT and IRT models in [19]. The evaluation is based on data from different intelligent tutoring systems.

In this work, we are interested in the properties of hybrid approaches combining latent factor and knowledge tracing models. In extension to previous work and especially to [19], we empirically evaluate the performance characteristics of the two recent hybrid models LFKT and FAST on synthetic data and compare them to the underlying approaches of BKT and IRT. We sample from a combined student model that encompasses all four models. By using synthetic data generated from the combined model, we show the robustness of the models under breaking model assumptions. By evaluating the models on 66'000 different parameter configurations we are able to rigorously explore the parameter space to demonstrate the relative performance gain between models for various regions of the parameter space. Our findings show that for the generated data sets FAST significantly outperforms all other methods for predicting the task outcome and that BKT is significantly better than FAST and LFKT at predicting the latent knowledge state. Furthermore we are able to identify the influence of different properties of a data set on model performance using regression and show best and worst case performances of the models.

2. INVESTIGATED MODELS

In an intelligent tutoring system a student is typically presented with a set of tasks to learn a specific skill. For each student n the system chooses at time t an item i from a set of items corresponding to a particular skill. The system then observes the answer $y_{n,t}$ of the student, which is assumed to be binary in this work. In the following, we briefly present four common techniques to model various latent states of the student and the tutoring environment.

BKT. Bayesian Knowledge Tracing (BKT) [9] models the knowledge acquisition of a single skill and is a special case of a Hidden Markov Model (HMM) [29]. BKT uses two latent states (*known* and *unknown*) to model if a student n has mastered a particular skill $k_{n,t}$ at time t, and two observable states (*correct* and *incorrect*) to represent the outcome of a particular task. Therefore, the probabilistic model can be fully described by a set of five probabilities. The initial probability of knowing a skill a-priori $p(k_{n,0})$ is denoted by p_I. The transition from one knowledge state $k_{n,t-1}$ to the next state $k_{n,t}$ is described by the probability p_L of transitioning from the *unknown* latent state to the *known* state and the probability p_F of transitioning from the *known* to the *unknown* state:

$$p(k_{n,t}) = k_{n,t-1}(1 - p_F) + (1 - k_{n,t-1})p_L. \quad (1)$$

In the case of BKT, p_F is fixed at 0. Finally, the task outcomes $y_{n,t}$ are modeled as

$$p(y_{n,t}) = k_{n,t}(1 - p_S) + (1 - k_{n,t})p_G, \quad (2)$$

where p_S denotes the *slip probability*, which is the probability of solving a task incorrectly despite knowing the skill, and p_G is the *guess probability*, which is the probability of correctly answering a task without having mastered the skill. Learning the parameters for a BKT model is done using maximum likelihood estimation (MLE).

IRT. Item Response Theory (IRT) [36] models the response of a student to an item as a function of latent student abilities θ_n and latent item difficulties d_i. The simplest form of an IRT model is the Rasch model, where each student n and each item i are treated independently. The outcome $y_{n,t}$ at time t is modeled using the logistic function

$$p(y_{n,t}) = \left(1 + e^{-(\theta_n - d_i)}\right)^{-1}. \quad (3)$$

A student with an ability of $\theta_n = d_i$ has a 50% chance of getting item i correct. In contrast to BKT, IRT does not model knowledge acquisition. The model parameters for the Rasch model are learned using EM.

LFKT. The Latent Factor Knowledge Tracing (LFKT) [18] model combines BKT and IRT using a hierarchical Bayesian model. On the basis of the BKT model, slip and guess probabilities are individualized based on student ability and item difficulty as

$$p_{G_{n,t}} = \left(1 + e^{-(d_i - \theta_n + \gamma_G)}\right)^{-1} \quad (4)$$

$$p_{S_{n,t}} = \left(1 + e^{-(\theta_n - d_i + \gamma_S)}\right)^{-1}, \quad (5)$$

where γ_G and γ_S are offsets for the guess and slip probabilities. The model is fit by calculating Bayesian parameter posteriors using Markov Chain Monte Carlo.

FAST. Feature Aware Student Knowledge Tracing (FAST) [14] allows for unification of BKT and IRT as well, but generalizes the individualized slip and guess probabilities to arbitrary features. Given a vector of features $\mathbf{f}_{n,t}$ for a student n at time t the adapted emission probability reads as

$$p(y_{n,t}) = \left(1 + e^{-(\boldsymbol{\omega}^T \mathbf{f}_{n,t})}\right)^{-1}, \quad (6)$$

where $\boldsymbol{\omega}$ is a vector of learned feature weights. If a set of binary indicator functions for the items and the students are used, FAST is able to represent the item difficulties d_i and student abilities θ_n from the IRT model. The parameters are fit using a variant of EM [6].

3. SYNTHETIC DATA GENERATION

Synthetic data is needed to have ground truth about the underlying data generating model, which enables the experimental evaluation of various properties of a model.

The sampling procedure starts by generating N student abilities θ_n from a normal distribution $N(0, \sigma)$. Then, it generates I item difficulties d_i from a uniform distribution $U(-\delta, \delta)$. Based on the initial probability p_I and the learn probability p_L a sequence of knowledge states $k_{n,0}, k_{n,1}, \ldots, k_{n,T}$ is sampled based on (1) and we therefore simulate data from only one skill. The time t^* at which $k_{n,t^*} = 1$ for the first time is considered as the moment of mastery. The number of sampled knowledge states is then given as $T = t^* + L$, where L denotes the lag of the simulated mastery learning system. For each student we generate a random sequence of items, i.e., item indices i. Arbitrary features from the training environment, such as answer times, help calls, problem solving strategy, engagement state of the student and gaming attempts, can have an influence on the performance of a student. To simulate those influences in a principled way, a single feature f is added to the data generating model with a varying feature weight ω (and thus varying correlation to the task outcomes $y_{n,t}$).

Based on these quantities, we sample the observations $y_{n,t}$ from a Bernoulli distribution with probability

$$p(y_{n,t}) = \left(1 + e^{-(\theta_n - d_i - \log \gamma_{n,t} + \omega f_{n,t})}\right)^{-1}, \qquad (7)$$

where

$$\gamma_{n,t} = (k_{n,t}(1 - p_S) + (1 - k_{n,t})p_G)^{-1} - 1.$$

Figure 1 gives a graphical overview of the described sampling procedure. Our sampling model has the following nine parameters: $p_I, p_L, p_S, p_G, \delta, \sigma, \omega, I, N$. The described sampling procedure allows sampling of data that exactly matches the model assumptions of all four models. To sample BKT data we set $\delta = \sigma = \omega = 0$ and (7) simplifies to the standard BKT formulation. By setting $p_S = p_G = 0.5$ and $\omega = 0$ we can sample from an IRT model. To sample from an LFKT model we set $\omega = 0$ and for FAST none of the parameters are restricted.

4. EXPERIMENTAL SETUP

Parameter space. We generated a vast number of parameter configurations in order to analyze the four models. The set of parameter configurations has been carefully designed to match real world conditions. The BKT parameters (p_I, p_G, p_S, p_L) are based on the parameter clusters found on real world data [30]. Using a normal distribution with a standard deviation of 0.02, we sampled up to 30 points (depending on the cluster size) around each cluster mean. According to common practice [16] we scaled the student abilities θ_n to have a mean of 0 and a variance of 1 and therefore $\sigma = 1$. We sampled the parameter δ (determining the range of the item difficulties) uniformly from $[0, 3]$ (according to [16]). Despite simulating only one skill, we varied the item difficulties to account for the fact that skill models tend to be imperfect in practice [7, 32, 20]. In accordance to the item difficulties, the feature weight ω was varied uniformly across $[0, 1.5]$. Feature values $f_{n,t}$ were sampled from the uniform distribution $U(-1, 1)$.

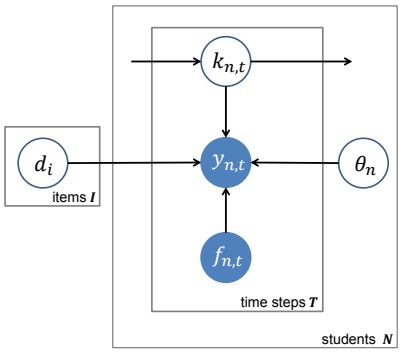


Figure 1: Combined student model used for synthetic data generation. The model corresponds to LFKT with the addition of a single feature. The relative dependencies of the observable nodes (blue) and the latent nodes (white) are shown. $k_{n,s}$ denotes the latent knowledge state, d_i the item difficulty, θ_n the student ability, $y_{n,t}$ the observation, and $f_{n,t}$ the feature value.

For every parameter configuration we generated five folds with $N = 300$ simulated students. Each fold was randomly split up into two parts of equal number of students. The first part was used as training data and the second part for testing. Therefore, the training data did contain unseen students only. As we simulated data from a mastery learning environment the number of tasks simulated for each student was determined by the moment of mastery. Based on the results presented by [11], we set the lag of the simulated system to $L = 4$ tasks from the moment of mastery. We simulated $I = 15$ different items with random item order.

In total, we generated 66′000 parameter configurations for $p_I, p_G, p_S, p_L, \delta, \omega$, this amounts total evaluation time (training and test) of 1'280 hours and 1'351 hours for LFKT and FAST respectively. The evaluation time for the BKT was 99 minutes and all configurations were evaluated in 58 minutes for the IRT model.

Implementation. To train BKT models we used our custom code that trains BKT using the Nelder-Mead simplex algorithm minimizing the log-likelihood. We thoroughly tested our implementation against the BKT implementation of [39]. The IRT models were fit by joint maximum likelihood estimation [21] implemented in the psychometrics library[1]. FAST using IRT features was shown to be equivalent to LFKT except for the parameter estimation procedure [19]. As this work did not investigate different parameter estimation techniques, both models were trained and evaluated using the publicly available FAST student modeling toolkit[2].

5. RESULTS AND DISCUSSION

Using the generated data, we investigated the performance characteristics of the four models and evaluated their predictive power and robustness under varying parameter configurations. For our results we generated 66′000 parameter

[1]An open source Java library for measurement, available at https://github.com/meyerjp3/psychometrics.
[2]http://ml-smores.github.io/fast/

configurations, and for each of them we generated synthetic data for $1'500$ students. Note that there are many ways to characterize performance differences among student models and we only cover a subset of these possibilities.

5.1 Error Metrics

The right choice of error metrics when evaluating student models has recently gained increased interest in the EDM community. In [28] some of the common error metric choices are discussed, highlighting possible issues with the accuracy and area under the ROC curve (AUC) measure. Correlations between various performance metrics and the accuracy of predicting the moment of mastering a skill has been investigated in [26], showing that the F-measure (equaling to the harmonic mean of precision and recall) and the recall are two metrics with a high correlation to the accuracy of knowledge estimation. The root mean squared error (RMSE) and log-likelihood, on the other hand, are well suited if one wants to recover the true learning parameters. Similarly, [10] concluded from results of 26 synthetic data sets that RMSE is better at fitting parameters than the log-likelihood.

In line with this previous work we investigated correlations between accuracy, RMSE and F-measure across all four models. For this, all models were trained and evaluated on data using $66'000$ different parameter configurations. All metrics are strongly correlated $|\rho| > 0.75, p \ll 0.001$. Our inspections of the metric correlations revealed no significant differences in the metric correlations among the different models. Thus, to a large extent the measures capture equal characteristics for the models we considered in this work. In the following, we therefore focus our analysis on the RMSE measure.

5.2 Model Comparison

Overall Performance. In a first step we investigated the overall performance of the models. For every parameter configuration, we calculated the average RMSE over the five generated folds. Table 1 summarizes the parameters for the best and worst data set for every model when model assumptions are met (see Section 3). Results show that all models that model a knowledge state (all except IRT) perform best if the slip probability is low and the guess probability is high. This leads to a data set that exhibits a high ratio of correct observations. IRT performs best on data that has very distinguished item difficulties (δ is high). Notably the best performance of FAST is achieved on a data set without features ($\omega = 0$). We assume that this is due to the decreased complexity of the data set, compared to one that exhibits high ω. Consistently, worst case data sets exhibit high symmetric values for guess and slip probabilities. In the case of LFTK and FAST worst case data sets additionally do not distinguish between items (difficulty range $\delta = 0$) and for FAST the feature weights are low.

We then performed the non-parametric Friedman test over all parameter configurations to assess performance differences between the models. We found that there is a statistically significant difference in the performance of the models ($\chi^2(3) = 13'065, p < 0.0001$). Performing a post-hoc analysis using Scheffe's S procedure [31] shows all model differences to be significant at $p < 0.0001$ with mean ranks of 1.7156, 2.3017, 2.6898 and 3.2929 for FAST, LFKT, BKT, and IRT, respectively. FAST therefore significantly outperforms the other methods on our synthetic data sets. In [19] IRT performed not significantly worse than LFKT and FAST on four different data sets. The good performance of IRT was attributed to the deterministic item ordering that allows IRT to infer knowledge acquisition confounded with item difficulty. Our results support this hypothesis as in our synthetic data set the items are in random order and IRT exhibits the worst overall performance.

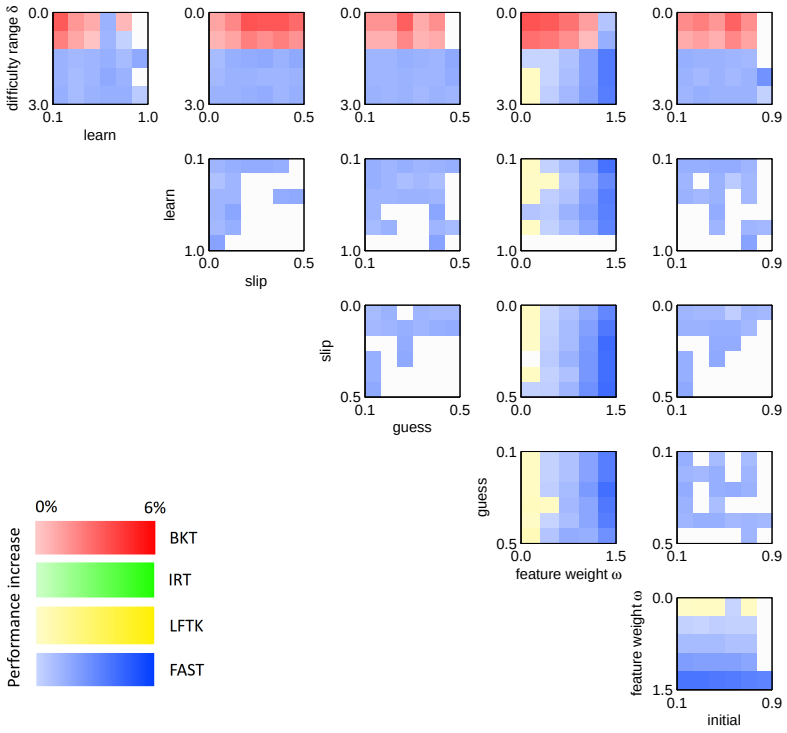
Table 1: Parameters of best and worst case data sets for each model. We only considered data sets that meet the model assumptions. Parameters denoted with * are fixed according to the model assumptions (see Section 3).

Model	δ	pI	pL	pS	pG	ω	RMSE
BKT							
Best	0.00*	0.71	0.41	0.01	0.47	0.00*	0.25
Worst	0.00*	0.10	0.12	0.50	0.49	0.00*	0.48
IRT							
Best	3.00	0.10	0.08	0.50*	0.50*	0.00*	0.42
Worst	0.00	0.10	0.10	0.50*	0.50*	0.00*	0.50
LFKT							
Best	0.75	0.69	0.40	0.01	0.46	0.00*	0.25
Worst	0.00	0.53	0.16	0.28	0.29	0.00*	0.51
FAST							
Best	0.75	0.67	0.40	0.01	0.46	0.00	0.25
Worst	0.00	0.56	0.16	0.28	0.28	0.00	0.51

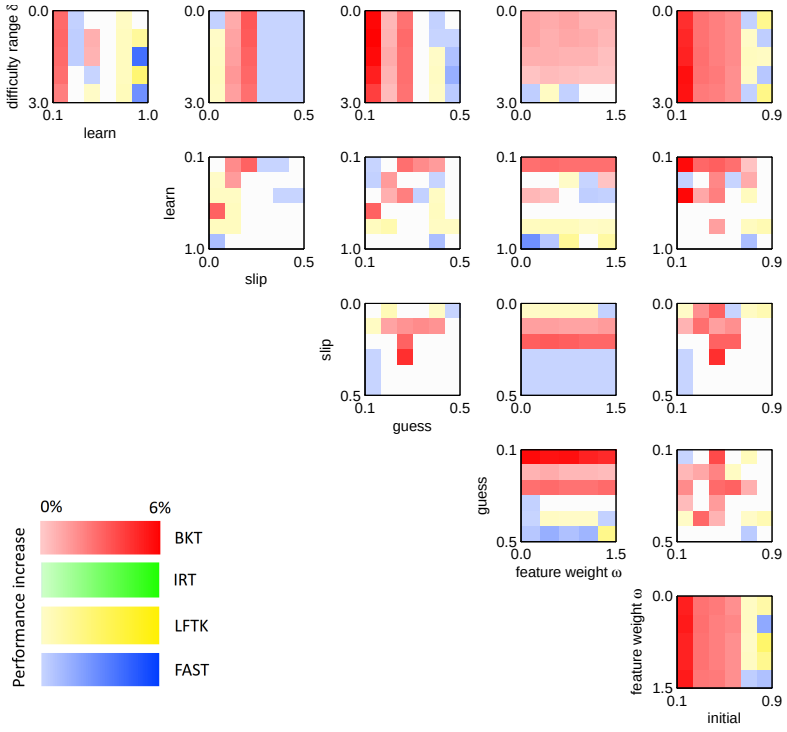
Parameter Space Investigation. To gain a better understanding of the performance characteristics of the different models, we analyzed their performances across the parameter space. For every pair of parameters p_i and p_j, we divided the parameter configurations into bins with similar values for p_i and p_j. We used five bins for each parameter (p_i and p_j) resulting in a total of 25 bins. Performance of each model was assessed by calculating the mean RMSE for each bin. Significance of the observed performance differences was computed using the Friedman test and $p < 0.05$.

Figure 2a shows the relative performance of the best model for each parameter pair. The models are color-coded: BKT is shown in red, IRT in green, LFKT in yellow, and FAST in blue. The color gradient indicates the relative improvement of the winning model over the second best model, where darker colors indicate higher values. White-colored areas indicate that there is no significant difference between the models. The plot shows that FAST is robust to parameter variations and outperforms the other models in large parts of the parameter space. In parts with low feature weights, i.e., where the feature f shows only a low correlation with task outcomes, LFKT outperforms FAST. When the variance δ of item difficulties d_i is low, BKT is the best model. A low variance in d_i implies a good skill model, with all tasks having approximately the same difficulty.

In contrast to Figure 2a, where we assessed the prediction



(a) Relative improvement in task outcome prediction (RMSE).



(b) Relative improvement in knowledge state prediction (RMSE).

Figure 2: Best performing models (RMSE) regarding prediction of task outcomes (a) and knowledge state prediction (b). The color for each bin indicates the best performing model, averaged over all other parameters. We investigated BKT (red), IRT (green), LFKT(yellow), and FAST(blue). White-colored bins exhibit no significant difference in model performance. The color brightness indicates the relative improvement of the best performing model over competing models, with dark colors referring to higher values. FAST is robust to parameter variations and outperforms the other models in large parts of the parameter space when predicting task outcomes (a). BKT is the best model if the variance of the item difficulty is low (a). BKT is superior to the other models in large parts of the parameter space when predicting knowledge states (b).

of task outcomes, we analyzed the quality of the prediction of knowledge states $k_{n,t}$ using the RMSE in Figure 2b. Ultimately, we want to predict whether a student has mastered a skill or not [26, 3]. The plot uses the same parameter pairs and color codings as Figure 2a. Interestingly, LFKT and FAST are not superior to BKT when it comes to prediction of the latent state. The additional parameters that LFKT and FAST use have a direct influence on the predicted task outcomes and therefore improve performance when predicting task outcomes. They have, however, no direct influence on the latent state $k_{n,t}$ of the model.

Robustness. Next, we tested the robustness of the different models against each other. We generated ideal data (meeting the model assumptions) for all the models and then interpolated the parameter values between these ideal cases. The classes of data sets that meet the model assumptions for the four models are described in Section 3. From every class of data sets, we selected the extreme case with the least amount of noise. In the following, we describe these cases.

For BKT, data is generated using $\delta = \omega = 0$, assuming a perfect skill model (all tasks with same difficulty) and setting the influence of additional (not captured) features to 0. Furthermore, we removed the randomness by setting $p_G = p_S = 0$. For IRT, the extreme case data was generated using $p_G, p_S = 0.5$, $\omega = 0$ and by additionally setting $\delta = 3$. As LFKT is a combination of IRT and BKT, we set the parameters to $p_G, p_S = 0.25$ and $\delta = 1.5$. Furthermore, we set $\omega = 0$, again assuming no influence of not captured features. For FAST we used the same parameters as for LFKT, but additionally introduced a feature influence by setting $\omega = 1.5$. We linearly interpolated the parameter space in-between these extreme cases to asses model robustness when model assumptions are violated. Figure 3 displays the model with best RMSE in this subspace that contains the extreme (ideal) cases, where p_L and p_I are averaged over the BKT parameter clusters presented in [30]. From these results, we can see that BKT tends to be robust to increased feature influence as long as $p_G, p_S \leq 0.15$. If the feature weight $\omega > 0.75$, FAST outperforms all the other classifiers. For large differences in item difficulties and large guess and slip probabilities, LFKT has a slight advantage over IRT.

5.3 Parameter Influence

To analyze the influence of the model parameters on the performance of the student models, we used linear regression to predict the RMSE based on the parameters of the sampling model. This allowed us to identify statistically significant correlations between the sampling parameters and the performance of the models despite the high dimensionality of the parameter space.

The sampling parameters have a direct influence on the ratio of correct observations in the data, e.g., a high learning probability with low guess and slip parameters leads to a high ratio of correct observations. Further, if the parameters model fast learners then the average number of tasks tends to be low since we are simulating a mastery learning environment. The three models IRT, LFKT and FAST which explicitly model items are sensitive to this kind of lacking data, as by having fewer observed items per student the estimation of item difficulty becomes more difficult. To

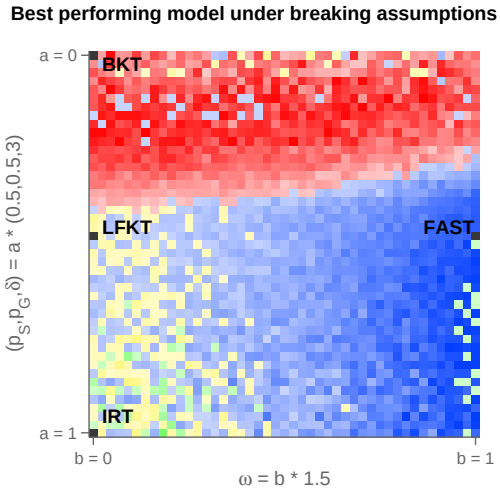


Figure 3: Relative model performance on ideal data sets generated by linearly interpolating between parameters. The colors refer to the models BKT (red), IRT (green), LFKT (yellow) and FAST (blue). The color gradient indicates the relative performance as in Figure 2a. BKT and FAST are more robust to the invalid assumptions of our experiment than IRT and LFKT.

investigate the effect of both factors, we added the two variables *correct ratio* and *average number of tasks* as predictors to the regression model. In order to make correlation coefficients comparable, all sampling parameters have been normalized to have mean 0 and standard deviation 1.

Figure 4 shows the regression coefficients for all four models, with red and green denoting statistically significant and not significant coefficients, respectively. The variables *correct ratio* and *average number of tasks* have the largest influence on the RMSE. Both effects are significant and positive (reducing the RMSE). A larger range of item difficulties δ has a positive influence on the performance of all models except for the BKT model. This is expected as BKT does not account for variations in item difficulty and thus larger variations in item difficulties are treated as noise by BKT, which makes prediction harder. IRT, LFKT and FAST, on the other hand, benefit from larger variations. We assume that this is due to the better identifiability of the effects of the different items. Interestingly, increasing the feature range ω has no significant negative effect for the models that do not take features into account (BKT, IRT, LFKT), but has a positive effect for FAST. The initial probability and the learning probability have a small negative and small positive effect on performance, respectively. While these coefficients are partially significant they have very small magnitude. The positive effect of the slip probability p_S for all models except BKT (the effect is not significant) is rather surprising. However, the effect of a high slip probability in our sampling model is that it weakens the influence of the latent knowledge state on the task outcomes. This could explain the positive influence for models that estimate item difficulty, since the difficulty estimates are less convoluted with effects from the knowledge state. Further work is needed to prove this effect.

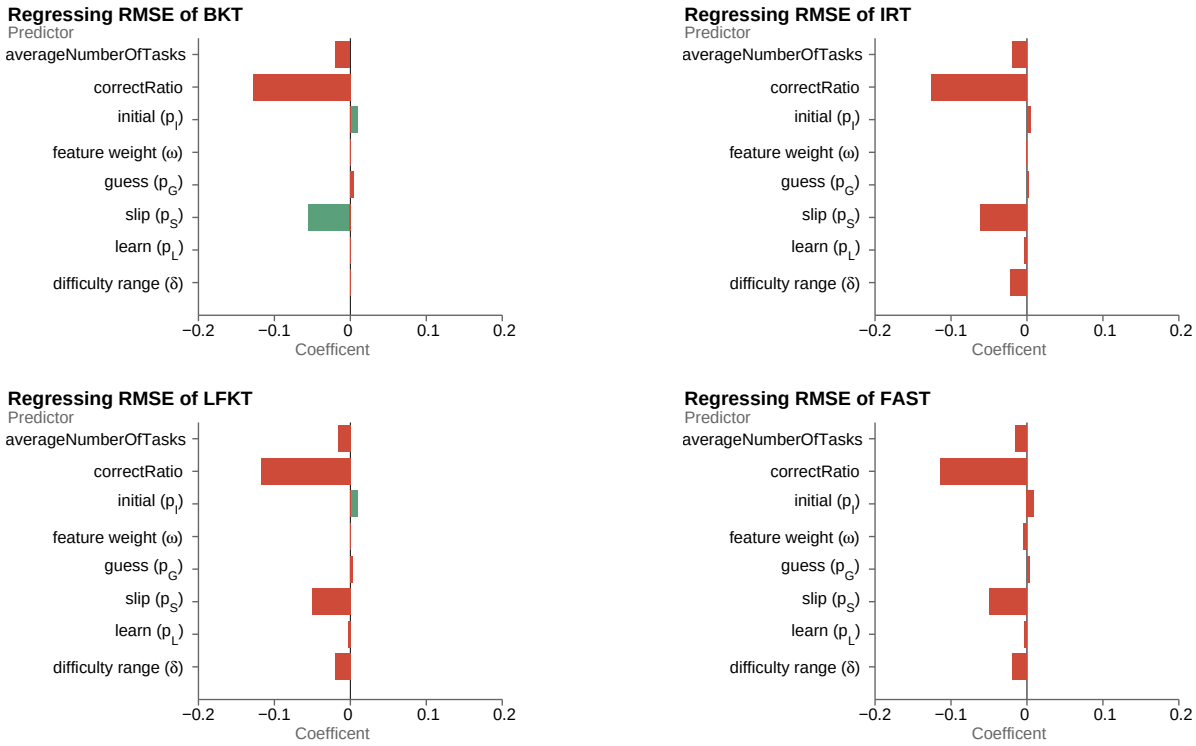


Figure 4: Regression coefficients to predict RMSE based on the sampling parameter values for the models BKT, IRT, LFKT and FAST. Parameters with positive coefficients have a negative effect on the performance and vice versa. Red denotes significant coefficients with $p < 0.001$, green coefficients are not significant.

6. CONCLUSIONS

In this work, we investigated the performance characteristics of latent factor and knowledge tracing models by exploring their parameter space. To do so, we generated a vast amount of 66'000 synthetic data sets for different parameter configurations containing data for 1'500 students each. Synthetic data allowed us to study the model performances under different parameter settings, and to test the robustness of the models against violations of specific model assumptions.

We showed best and worst case performances for all the models and investigated the relative performance gain in various regions of the parameter space. Our results showed that the two recently developed models LFKT and FAST, which synthesize item response theory and knowledge tracing, perform better than BKT and IRT. FAST even significantly outperformed LFKT if reasonable features can be extracted from the learning environment. Interestingly, IRT exhibited the worst performance, which supports the hypothesis by [19] that random item ordering has a negative influence on the performance of IRT models. However, more analyses are needed to investigate this effect thoroughly. Further, we investigated the models' abilities to predict the latent knowledge state and demonstrated that LFKT and FAST are outperformed by BKT. This raises the question of how to adjust the two recent methods LFKT and FAST if the aim is to predict knowledge states; we leave this exploration for future work. The analysis of the model robustness revealed that BKT is robust to increased feature influence for small guess and slip probabilities. For larger guess and slip, FAST outperformed the other methods.

While all sampling parameters have been carefully chosen to match real world conditions, we expect real world data to exhibit more noise and additional effects not covered by our synthetic data. Thus, the achieved performance can be considered an upper bound on the performance achievable in real world settings. The performance of BKT depends on the quality of the underlying skill model. We have simulated imperfect skill models by introducing item effects, but we did not take other sources for imperfect skill models into account. Furthermore, the simulated data consisted of a fixed set of items. For tutoring systems offering many variations of tasks, reliable estimation of item effects is challenging, which in turn influences the performance of IRT, LFKT and FAST. Moreover, the performance of FAST is driven by feature quality, which may vary between different tutoring systems.

Finally, it remains questionable whether and how the performance of the investigated techniques influences the learning outcome of students in a tutoring system. We show relative improvements in RMSE between models of up to 6%. However, the effect of small-scale improvements in the accuracy of student models on the learning outcome has been discussed controversially [4, 39].

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

7. REFERENCES

[1] 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. ITS*, 2008.

[2] R. S. Baker, A. T. Corbett, S. M. Gowda, A. Z. Wagner, B. A. MacLaren, L. R. Kauffman, A. P. Mitchell, and S. Giguere. Contextual Slip and Prediction of Student Performance after Use of an Intelligent Tutor. In *Proc. UMAP*, 2010.

[3] R. S. Baker, A. B. Goldstein, and N. T. Heffernan. Detecting the moment of learning. In *Proc. ITS*, 2010.

[4] J. Beck and X. Xiong. Limits to accuracy: how well can we do at student modeling? In *Proc. EDM*, 2013.

[5] J. E. Beck and K. M. Chang. Identifiability: A fundamental problem of student modeling. In *Proc. UM*, 2007.

[6] T. Berg-Kirkpatrick, A. Bouchard-Côté, J. DeNero, and D. Klein. Painless unsupervised learning with features. In *Proc. NAACL-HLT*, 2010.

[7] 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. AIED*, 2007.

[8] H. Cen, K. R. Koedinger, and B. Junker. Comparing two IRT models for conjunctive skills. In *Proc. ITS*, 2008.

[9] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *UMUAI*, 1994.

[10] A. Dhanani, S. Y. Lee, P. Phothilimthana, and Z. Pardos. A comparison of error metrics for learning model parameters in Bayesian knowledge tracing. Technical report, UCB/EECS-2014-131, EECS Department, University of California, Berkeley, 2014.

[11] S. Fancsali, T. Nixon, and S. Ritter. Optimal and worst-case performance of mastery learning assessment with Bayesian knowledge tracing. In *Proc. EDM*, 2013.

[12] G. H. Fischer and I. W. Molenaar. *Rasch models: Foundations, recent developments, and applications*. Springer Science & Business Media, 1995.

[13] Y. Gong, J. E. Beck, and N. T. Heffernan. Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. In *Proc. ITS*, 2010.

[14] J. P. González-Brenes, Y. Huang, and P. Brusilovsky. General features in knowledge tracing to model multiple subskills, temporal item response theory, and expert knowledge. In *Proc. EDM*, 2014.

[15] J. Gu, H. Cai, and J. E. Beck. Investigate performance of expected maximization on the knowledge tracing model. In *Proc. ITS*, 2014.

[16] D. Harris. Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice*, 1989.

[17] J. Johns and B. Woolf. A dynamic mixture model to detect student motivation and proficiency. In *Proc. Artificial intelligence*, 2006.

[18] M. Khajah, R. Wing, R. Lindsey, and M. Mozer. Integrating latent-factor and knowledge-tracing models to predict individual differences in learning. In *Proc. EDM*, 2014.

[19] M. M. Khajah, Y. Huang, J. P. González-Brenes, M. C. Mozer, and P. Brusilovsky. Integrating knowledge tracing and item response theory: A tale of two frameworks. *Personalization Approaches in Learning Environments*, 2014.

[20] K. Koedinger, J. Stamper, E. McLaughlin, and T. Nixon. Using data-driven discovery of better student models to improve student learning. In *Proc. AIED*, 2013.

[21] J. Meyer and E. Hailey. A study of Rasch, partial credit, and rating scale model parameter recovery in WINSTEPS and jMetrik. *Journal of Applied Measurement*, 2011.

[22] Z. A. Pardos and N. Heffernan. Introducing item difficulty to the knowledge tracing model. In *Proc. UMAP*, 2011.

[23] Z. A. Pardos and N. T. Heffernan. Modeling individualization in a Bayesian networks implementation of knowledge tracing. In *Proc. UMAP*, 2010.

[24] 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. EDM*, 2010.

[25] Z. A. Pardos, S. Trivedi, N. T. Heffernan, and G. N. Sárközy. Clustered knowledge tracing. In *Proc. ITS*, 2012.

[26] Z. A. Pardos and M. Yudelson. Towards moment of learning accuracy. In *AIED Workshops*, 2013.

[27] P. I. Pavlik, H. Cen, and K. R. Koedinger. Performance factors analysis - a new alternative to knowledge tracing. In *Proc. AIED*, 2009.

[28] R. Pelánek. A brief overview of metrics for evaluation of student models. In *Approaching Twenty Years of Knowledge Tracing Workshop*, 2014.

[29] J. Reye. Student modelling based on belief networks. *IJAIED*, 2004.

[30] S. Ritter, T. K. Harris, T. Nixon, D. Dickison, R. C. Murray, and B. Towle. Reducing the knowledge tracing space. In *Proc. EDM*, 2009.

[31] H. Scheffe. *The analysis of variance*, volume 72. John Wiley & Sons, 1999.

[32] J. C. Stamper and K. R. Koedinger. Human-machine student model discovery and improvement using datashop. In *Proc. AIED*, 2011.

[33] B. van de Sande. Properties of the Bayesian knowledge tracing model. *JEDM*, 2013.

[34] Y. Wang and J. Beck. Class vs. student in a Bayesian network student model. In *Proc. AIED*, 2013.

[35] Y. Wang and N. T. Heffernan. The student skill model. In *Proc. ITS*, 2012.

[36] M. Wilson and P. De Boeck. Descriptive and explanatory item response models. 2004.

[37] Y. Xu and J. Mostow. Using logistic regression to trace multiple subskills in a dynamic Bayes net. In *Proc. EDM*, 2011.

[38] Y. Xu and J. Mostow. Using item response theory to refine knowledge tracing. In *Proc. EDM*, 2013.

[39] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon. Individualized Bayesian knowledge tracing models. In *Proc. AIED*, 2013.