Different parameters - same prediction: An analysis of learning curves

Tanja Käser
Departement of Computer Science
ETH Zurich
kaesert@inf.ethz.ch

Kenneth R. Koedinger
Human-Computer Interaction Insitute
Carnegie Mellon
koedinger@cmu.edu

Markus Gross
Departement of Computer Science
ETH Zurich
grossm@inf.ethz.ch

ABSTRACT

Using data from student use of educational technologies to evaluate and improve cognitive models of learners is now a common approach in EDM. Such naturally occurring data poses modeling challenges when non-random factors drive what data is collected. Prior work began to explore the potential parameter estimate biases that may result from data from tutoring systems that employ a mastery learning mechanism whereby poorer students get assigned tasks that better students do not. We extend that work both by exploring a wider set of modeling techniques and by using a data set with additional observations of longer-term retention that provide a check on whether judged mastery is maintained. The data set at hand contains math learning data from children with and without developmental dyscalculia. We test variations on logistic regression, including the Additive Factors Model and others explicitly designed to adjust for mastery-based data, as well as Bayesian Knowledge Tracing (BKT). We find these models produce similar prediction accuracy (though BKT is worse), but have different parameter estimation patterns. We discuss implications for use and interpretation of these different models.

Keywords

learning curves, logistic regression models, knowledge tracing, parameter fitting, prediction accuracy

1. INTRODUCTION

Modeling student knowledge is a fundamental task when working with intelligent tutoring systems. The selection of tasks and actions is based on the student model, therefore an accurate prediction of student knowledge is essential. The accuracy of the student model depends on the quality of the parameter fit. Parameter fitting is, however, not only important for prediction accuracy; the parameters of a model also contain information on how students learn.

A variety of approaches to assess, interpret and predict student knowledge have been proposed. Popular techniques to model student learning include Bayesian Knowledge Tracing (BKT) [8], Bayesian networks [4, 9, 10], performance factors analysis [21] and Additive Factors Models (AFM) [5, 6].

BKT is one of the most popular approaches for student modeling. Prediction accuracy of the original BKT model has been improved using clustering approaches [20] or individualization techniques, such as learning student- and skill-specific parameters [16, 19, 24, 26] or modeling the parameters per school class [21].

The AFM is a generalized linear mixed model [2] applying a logistic regression. It is widely used to fit learning curves and to analyze and improve student learning. AFM helps identify flat or ill-fitting learning curves that indicate opportunities for tutor or model improvement. Consistently low error curves indicate opportunities to reallocate valuable student time [5]. Consistently high error curves with poor fit indicate a miss-specified skill model that can be improved [15, 23] and used to design better instruction [14]. However, when working with mastery learning data sets, averaging over students who have different initial knowledge states and learning rates may lead to learning curves which show little student learning. It has been shown [17] that disaggregating a learning curve into curves for different subpopulations or mastery-align the learning curves provides more accurate metrics for student learning. However, so far there exist no comparisons between the properties of the different models, such as the parameter fit. Furthermore, the models were also not validated regarding prediction accuracy.

In this work, we therefore extensively evaluate the properties and parameters of different logistic regression models when fitting learning curves to a mastery learning data set containing students with heterogeneous knowledge levels. We turn the suggestions of [17] for fitting learning curves in BKT into logistic regression models and also introduce a further alternative model to the AFM. The data set at hand was collected from an intelligent tutoring system for learning mathematics and includes log files from 64 children with developmental dyscalculia and 70 control children. Our findings show that similar regression models predict very different amounts of learning for the same data. Furthermore, we demonstrate that different parameter fits lead to the same prediction accuracy on unseen data. For further validation, we compare prediction accuracy of logistic regression models to that of BKT and analyze how well these models generalize to new students. Our results demonstrate that logistic regression models outperform BKT regarding prediction accuracy on unseen data.

2. METHOD

In the following, we first introduce different logistic regression models and their properties. We then give a short overview of BKT and finally explain the experimental setup.

2.1 Logistic regression models

Logistic regression models are used in Item Response Theory (IRT) [25] to model the response (correct/wrong) of a student to an item. IRT is based on the idea that the probability of a correct response to an item is a mathematical function of student and item parameters. The logistic regression models presented in the following are based on this concept.

Additive Factors Model (AFM). The AFM [5, 6] is a logistic regression model fitting a learning curve to the data. In a logistic regression model, the observations of the students follow a Bernoulli distribution. A Bernoulli distribution is a binomial distribution with $n = 1$. Letting $y_{pi} \in \{0, 1\}$ denote the response of student p on item i, we obtain $y_{pi} \sim binomial(1, \pi_{pi})$. The linear component π_{pi} of the AFM can then be formulated as follows:

$$\pi_{pi} = logit(\theta_p + \sum_k q_{ik} \cdot (\beta_k + \gamma_k \cdot T_{pk})), \tag{1}$$

with $\theta_p \sim \mathcal{N}(0, \sigma_\theta^2)$. The AFM is a generalized linear mixed model with a random effect θ_p for student proficiency and fixed effects β_k (difficulty) and γ_k (learning rate) for the skills k (knowledge components). The learning rate γ_k is constrained to be greater than or equal to zero for AFMs. q_{ik} is 1, if item i uses skill k and 0 otherwise. Finally, T_{pk} denotes the number of practice opportunities student p had at skill k. The AFM is related to the linear logistic test model (LLTM) [25] and the Rasch model [25]. When removing the third term ($\gamma_k \cdot T_{pk}$) of Equation 1, we obtain an LLTM. Additionally assuming a unique-step skill model (one skill per step) results in the Rasch model. The intuition of the AFM is that the probability of a student getting a step correct is proportional to the amount of required knowledge of the student θ_p, plus the difficulty of the involved skills β_k and the amount of learning gained from each practice opportunity γ_k.

Learning curves are averaged over many students. The AFM aligns the students by opportunity count. When applied to mastery learning data, it therefore suffers from student attrition with increasing numbers of opportunities. Well performing students need few opportunities to master a skill and thus only the weaker students remain in the analysis for higher opportunity counts. This student attrition can lead to an underestimation of the learning rates γ_k. In the following, we therefore introduce alternative logistic regression models that adjust for mastery-based data.

Learning Gain Model (LG). With the LG model, we introduce a new alternative to the AFM. The LG model avoids student attrition by aligning the students at their first sample (when they start the training) and at their last sample, i.e., when they end the training (independent of whether they mastered the skill or not). The linear component of this model is very similar to that of the AFM:

$$\pi_{pi} = logit(\theta_p + \sum_k q_{ik} \cdot (\beta_k + \gamma_k \cdot N_{pk})), \tag{2}$$

where $N_{pk} \in [0, 1]$ denotes the normalized opportunity count of student p at skill k, i.e., we normalize over all opportunities student p had at skill k during the training. Rather than measuring the amount learnt per opportunity, this model estimates the learning gain of the students over the course of the training.

Alternative logistic regression models. To adjust for mastery-based data, alternative ways to fitting the curves have been proposed [17] for BKT. In the following, we reformulate these suggestions and apply them to logistic regression models. The **Mastery-Aligned Model (MA)** can be formulated using Equation 1, but with a different definition of T_{pk}. For the MA model, we count backwards: T_{pk} is the number of opportunities student p had at skill k as seen from mastery. T_{pk} is 0 at mastery, 1 at one opportunity before mastery and so on. Thus, the MA model aligns students at mastery, which solves the problem of student attrition. A different way to deal with student attrition is to group students by the number of opportunities needed to first master a skill. The linear component of this **Disaggregated Model (DIS)** can be defined as follows:

$$\pi_{pi} = logit(\theta_p + \sum_{k,m} q_{ik} \cdot (\beta_{k,m} + \gamma_{k,m} \cdot T_{pk})), \tag{3}$$

where the difficulty $\beta_{k,m}$ and the learning rate $\gamma_{k,m}$ are fit by skill k and mastery group m. By combining the MA and the DIS models, the **Mastery-Aligned and Disaggregated Model (DISMA)** can be constructed. This model disaggregates students into groups based on the number of opportunities needed until mastering the skill and furthermore aligns the students at mastery.

All models presented are generalized linear mixed models (GLMM) as the linear predictor π_{pi} contains random effects (for the students) in addition to the fixed effects (for the skills). GLMMs are fit using maximum likelihood, which involves integration over the random effects [3]. Integration is performed using methods such as numeric quadrature or Markov Chain Monte Carlo.

2.2 Bayesian Knowledge Tracing

BKT [8] is a popular approach for modeling student knowledge. BKT models are a special case of Hidden Markov Models (HMM) [22]. In BKT, student knowledge is modeled by one HMM per skill (or knowledge component). The latent variable of the model represents the student knowledge. It indicates whether a student has mastered the skill in question and is therefore binary. The state of this variable is inferred by binary observations, i.e., correct or wrong answers to tasks associated with the skill in question. A HMM can be specified using five parameters. The transmission probabilities of the model are defined by the probability p_L of a skill transitioning from not known to known state and the probability p_F of forgetting a previously known skill. 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 define the emission probabilities of the model. And finally, p_0 denotes the probability of knowing a skill a-priori. In BKT, the forget probability p_F is assumed to be 0 and therefore a BKT model can be specified with the four parameters $\theta = \{p_0, p_L, p_s, p_g\}$.

An important task when working with BKT models is parameter learning. The learning task can be formulated as follows: Given a sequence of student observations $\mathbf{y} = \{y_t\}$ with $t \in [1, T]$, what are the parameters $\theta = \{p_0, p_L, p_s, p_g\}$ that maximize the likelihood of the data $p(\mathbf{y}|\theta)$. BKT models have been fit using expectation maximization [7], brute-force grid search [1] or gradient descent [26].

2.3 Experimental setup

The training environment we use in this work consists of `Calcularis`, a tutoring system for children with difficulties in learning mathematics [11]. The program transforms current neuro-cognitive findings into the design of different instructional games, which are classified into two parts. The first part focuses on the training of different number representations and number understanding. In the second part, addition and subtraction are trained at different difficulty levels. Task difficulty depends on the magnitude of numbers involved, the complexity of the task and the means allowed to solve the task. The employed student model is a dynamic Bayesian network modeling different mathematical skills and their dependencies. The controller acting on the skill net is rule-based and allows forward and backward movements (increase and decrease of difficulty levels) [12, 13].

The data set used for the experimental evaluation was collected in a large-scale user study in Switzerland and Germany with 134 participants (69% females). 64 participants (73% females) were diagnosed with developmental dyscalculia (DD) and 70 participants (66% females) were control children (CC). All children were German-speaking and visited the 2^{nd}-5^{th} grade of elementary school (mean age: 8.68 (SD 0.84)). Children trained with the program for six weeks with a frequency of five times per week during sessions of 20 minutes. The collected log files contain at least 24 complete sessions per child. On average, each child solved 1521 tasks (SD 269) during the training. Results of the external pre- and post-tests demonstrated a significant improvement in spatial number representation, addition and subtraction after the training [11].

We investigated 20 addition and subtraction skills in the number range $0 - 100$. For our analyses, we used two versions of the data set. The first version (denoted as *Version 1* in the following) contains the samples of all children at the respective skills, while the second version (denoted as *Version 2* in the following) includes only children that mastered the respective skills. *Version 2* of the data set makes the inclusion of the MA and DISMA models possible. However, it excludes students not mastering a skill from the analysis, which leads to a more homogeneous, but due to the drop-out of many children with DD, also less interesting data set. *Version 1* of the data set contains $36'350$ solved tasks, while *Version 2* consists of $20'784$ tasks. External paper-pencil and computer-based arithmetic tests conducted at the beginning and at the end of the study demonstrated significant improvement in addition and subtraction in the number range $0 - 100$.

3. EVALUATION AND RESULTS

In a first study, we analyzed the parameter fit of different regression models and evaluated their performance in prediction of new items. Furthermore, we compared prediction accuracy of regression models to that of traditional BKT. We used all the samples until the children mastered a skill and predicted the outcome of the first re-test. In a second experiment, we evaluated the prediction accuracy of regression models as well as BKT when generalizing to new students. We fitted the model based on a subset of students and predicted the outcome for the rest of the students. Prediction accuracy for both experiments was measured using the root mean squared error (RMSE), the accuracy (number of correctly predicted student successes/failures based on a threshold of 0.5) and the area under the ROC curve (AUC). Prediction accuracy was computed using bootstrap aggregation with re-sampling ($n = 200$) in the first experiment and a student-stratified 10-fold cross validation in the second experiment.
Fitting for the regression models was done in `R` using the `lme4` package. To be able to compare the parameter fit of the different models, we did not constrain γ_k to be greater than or equal to zero. Parameters for BKT were estimated by maximizing the likelihood $p(\mathbf{y}|\theta)$ using a Nelder Mead simplex optimization [18]. This minimization technique does not require the computation of gradients and is for example available in `fminsearch` of `Matlab`. The following constraints were imposed on the parameters: $p_g \leq 0.3$ and $p_s \leq 0.3$.

3.1 Analysis of parameter fit

In this experiment, we investigated the parameter fit of three regression models on the data set *Version 1*: The AFM, the LG model and the DIS model. The three models obtain very different parameter estimations for the same data. While the AFM model predicts learning (positive γ_k) for 50% of the skills, the LG model fits positive learning rates γ_k for all skills and the DIS model obtains positive learning rates $\gamma_{k,m}$ for 92% of the cases. We therefore analyze the residuals and prediction accuracy of the different models in the following.

Residual analyses. All three models tend to overestimate the outcome for badly performing students and underestimate the outcome for well performing students. This finding is also visible in Fig. 1, which displays the mean residuals r with $r =$ *fitted outcome - true outcome* by estimated student proficiency θ_p. Furthermore, the residuals r are strongly correlated to student proficiency ($\rho_{AFM} = -0.9621$, $\rho_{LG} = -0.9612$, $\rho_{DIS} = -0.9532$). These results are as expected, because the models' predictions are averaged over all the students. While the residuals r are very similar for the AFM and the LG models, the DIS model exhibits less variance in student proficiency. As the students are grouped by the number of opportunities needed to master a skill, student proficiency within a group is more homogeneous.

For the AFM and the LG model, we also analyzed the mean residuals r regarding the skill parameters β_k and γ_k from the models. There are no significant correlations between

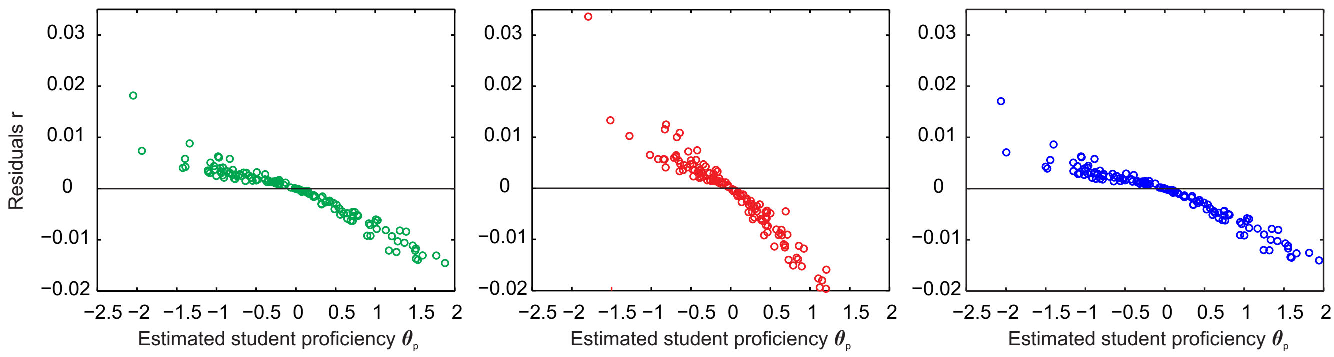

Figure 1: Mean residuals r by estimated student proficiency θ_p for the AFM (left), the DIS (middle) and the LG (right) model.

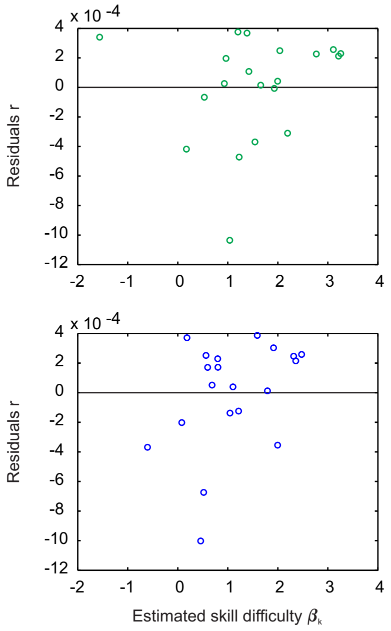

Figure 2: Mean residuals r by estimated skill difficulty β_k for the AFM (top) and the LG model (bottom).

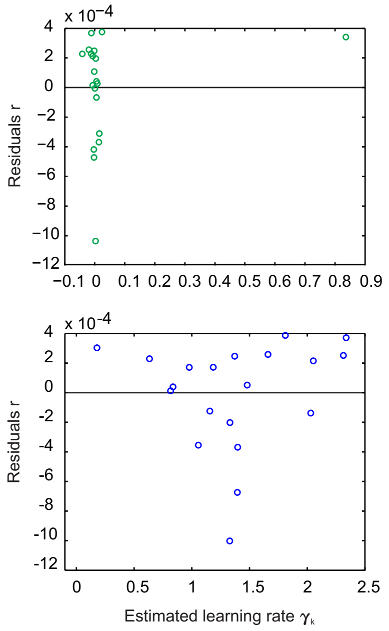

Figure 3: Mean residuals r by estimated learning rates γ_k for the AFM (top) and the LG model (bottom).

skill difficulty β_k and mean residuals r neither for the AFM ($\rho_{AFM} = 0.1677, p_{AFM} = .4798$) nor for the LG model ($\rho_{LG} = 0.3777$, $p_{AFM} = .1066$). From Fig. 2, which displays the mean residuals r by estimated skill difficulty β_k, it is also obvious that these measures are not related for both models. The residuals r are also not correlated to the estimated learning rate γ_k ($\rho_{AFM} = 0.2058$, $p_{AFM} = .3840$; $\rho_{LG} = 0.1051$, $p_{LG} = .6592$) as displayed in Fig. 3. Figure 3 demonstrates how different the parameter fits of the two models are regarding the learning rates γ_k. The AFM fits learning rates γ_k in a very small range around 0 and 45% of the learning rates are not significantly different from zero. The outlier stems from a skill played by only two students resulting in a total of 14 solved tasks. Learning rates γ_k fitted by the LG model are all positive and exhibit a larger variance. This larger variance appears to result from AFM having a bias to underestimate learning rate (because mastery leaves more poor students contributing to high opportunity counts) and LG having a bias to overestimate learning rate (because the adjusted end-point of all learning curves, the last opportunity that achieves mastery, is always successful whether or not it is a true or false positive).

The mean residuals r over time are displayed in Fig. 4. For the AFM and the DIS model, an averaging window ($n = 10$) was used to compute the mean residuals r with increasing opportunity count. Both models underestimate the outcome for less than 20 opportunities and overestimate it for larger numbers. For the AFM, this observation is confirmed by the significant positive correlation between the opportunity

Table 1: Prediction accuracy of first re-test for data set *Version 1* and *2*. The values in brackets denote the standard deviations. The best model per error measure is marked (*).

		RMSE	Accuracy	AUC
Data set: Version 1	AFM	0.3562 (0.0101)*	0.8391 (0.0119)	0.6825 (0.0230)*
	LG	0.3587 (0.0125)	0.8451 (0.0113)*	0.6778 (0.0250)
	DIS	0.3780 (0.0140)	0.8394 (0.0122)	0.6054 (0.0255)
	BKT	0.3614 (0.0111)	0.8428 (0.0118)	0.6033 (0.0250)
Data set: Version 2	AFM	0.3563 (0.0114)*	0.8474 (0.0123)*	0.6622 (0.0250)*
	LG	0.3666 (0.0124)	0.8416 (0.0107)	0.6602 (0.0245)
	DIS	0.3765 (0.0141)	0.8416 (0.0120)	0.5998 (0.0290)
	MA	0.3633 (0.0117)	0.8401 (0.0114)	0.6508 (0.0255)
	DISMA	0.3783 (0.0133)	0.8396 (0.0116)	0.6011 (0.0256)
	BKT	0.3613 (0.0111)	0.8423 (0.0115)	0.6102 (0.0302)

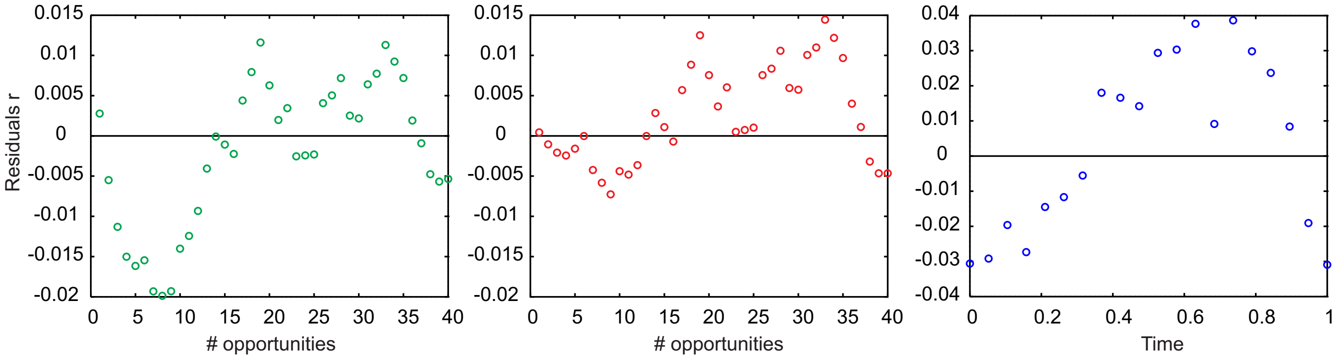

Figure 4: Mean residuals r by opportunity count for the AFM (left) and the DIS (middle) model and by normalized opportunity count for the LG (right) model.

count and the mean residuals r ($\rho_{AFM} = 0.3950$, $p_{AFM} < .001$). This result probably stems from the fact that the well performing students master the skills much faster and therefore student numbers drop with higher opportunity counts. The DIS model exhibits a lower variance, as this model groups the students by the number of opportunities needed to master a skill and thus student performance within a group is more homogeneous ($\rho_{DIS} = 0.0860$, $p_{DIS} = .4785$). For the LG model, the mean residuals r are plotted by the normalized opportunity count in Fig. 4 (right). The LG model underestimates the outcome in the beginning and in the end and overestimates in-between. Through normalizing the opportunity count, we align the beginning and the end of the training for each student. We therefore end up with more observations from low performing students in the middle and the model overestimates the outcome in this part.

Re-test prediction. The residual analyses demonstrate that the models interpret the same data very differently, i.e., the parameter fit and properties of the models vary a lot. To validate these different parameter fits, we computed the prediction accuracy for the first re-test (data set *Version 1*) and compared it to a BKT model. The observed mean outcome over all re-tests is high with 0.8419. The AFM underestimates the true outcome with an average prediction of 0.8287, while the LG (average prediction 0.9108) and DIS models (average prediction 0.9488) overestimate the true outcome. Prediction accuracy for the different models is listed in Tab. 1. The AFM shows the best RMSE ($RMSE_{AFM} = 0.3562$) and AUC ($RMSE_{AUC} = 0.6825$), while the LG models exhibits the highest accuracy ($Accuracy_{LG} = 0.8451$). As the performance of students is generally high, RMSE and AUC are, however, better quality measures than accuracy. The LG model performs second best in RMSE ($RMSE_{LG} = 0.3587$) and AUC ($AUC_{LG} = 0.6778$). However, the small differences between the AFM and the LG model along with the high variances of the error measures indicate that there are no significant differences between the two models. The DIS model on the other hand demonstrates a considerably higher RMSE ($RMSE_{DIS} = 0.3780$) and also exhibits a low AUC ($AUC_{DIS} = 0.6054$) compared to the two other regression models. The DIS model estimates the parameters $\beta_{k,m}$ and $\gamma_{k,m}$ by skill and mastery group. The resulting large number of parameters produces overfitting. Performance on the training data set supports the overfitting hypothesis: The DIS model outperforms the AFM and the LG model in RMSE, accuracy and AUC on the training data set.

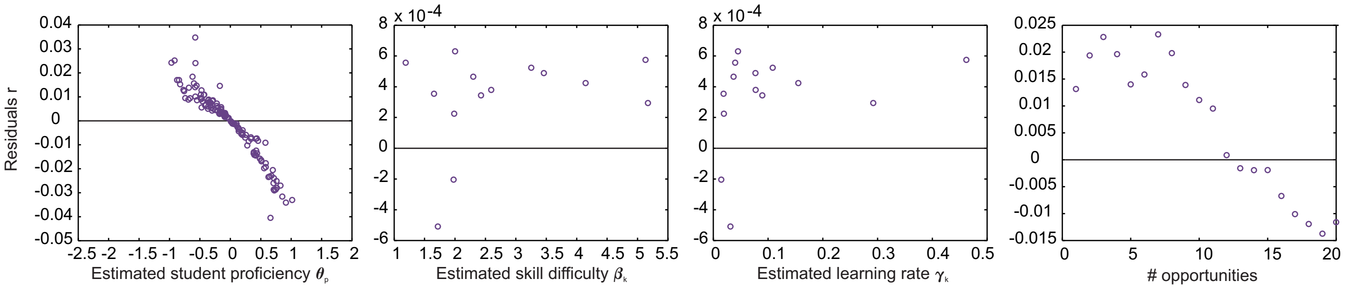

Figure 5: Mean residuals r by estimated student proficiency θ_p (left), skill difficulty β_k (center left), learning rates γ_k (center right) and opportunity count (right) for the MA model.

Interestingly, the AFM and the LG model also outperform the BKT model. The RMSE of BKT ($RMSE_{BKT} = 0.3614$) is higher than those of the two regression models, but standard deviations are again large. BKT exhibits especially a lower performance in AUC ($AUC_{BKT} = 0.6033$). The better performance of the regression models might come from two facts: First, the regression models fit the parameter θ_p for the individual student's proficiency, while traditional BKT does not do any student individualization. Second, BKT assumes that there is no forgetting, while the regression models are allowed to fit negative learning rates γ_k. However, the time between mastering a skill and the first re-test tends to be long. On average, the first re-test was done after 140 opportunities. A logistic regression analysis shows, that there is indeed a small, but significant amount of forgetting (intercept = 1.8545, slope = -0.0012) in the data. The probability of being correct at mastery amounts to 0.8647 and decreases to 0.8419 after 140 opportunities. Note, however, that the forgetting hypothesis is only valid for the AFM, as learning rates γ_k are all positive for the LG model.

Experiments on data set *Version 2*. To be able to include the MA and DISMA models in our analyses, we also evaluated prediction accuracy for the first re-test based on data set *Version 2*.
For this version of the data set, the LG and MA models predict positive learning rates γ_k for 100% of the skills, while the AFM fits positive learning rates γ_k for 54% of the skills. The DIS and DISMA models show positive learning rates $\gamma_{k,m}$ for 90% of the mastery groups. Residuals r of the DISMA model are very similar to those of the DIS model and we therefore only discuss the mean residuals r for the MA model. Figure 5 displays the mean residuals r by estimated student proficiency θ_p (left), skill difficulty β_k (center left), learning rates γ_k (center right) and over time (right). Similarly to the other models, the MA model tends to overestimate the well performing students and underestimate the weaker students (see Fig. 5 (left)). The correlation between estimated student proficiency θ_p and mean residuals r is again strong ($\rho_{MA} = -0.9497$, $p_{MA} < .001$). As for the other models, mean residuals r are uncorrelated to skill difficulty β_k ($\rho_{MA} = 0.2916$, $p_{MA} = .3118$) and to learning rates γ_k ($\rho_{MA} = -0.2993$, $p_{MA} = .2986$). The MA model fits positive learning rates γ_k for all skills k (see Fig. 5 (center right)). To compute the mean residuals r by opportunity count, we again used an averaging window ($n = 10$). Unlike the other models, the MA model overestimates the outcome in the beginning and underestimates it with increasing opportunity count. This result is due to the mastery alignment of the model: As well performing students need less opportunities to master a skill, student attrition occurs in the beginning, where only weaker students remain in the analysis.

We again validated the parameter fit of the different models by predicting the first re-test and comparing prediction accuracy to BKT. Prediction accuracy for the different models is listed in Tab. 1. The AFM performs best for all error measures ($RMSE_{AFM} = 0.3563$, $AUC_{AFM} = 0.6622$). The performance of the LG model ($RMSE_{LG} = 0.3666$, $AUC_{LG} = 0.6602$) is again very close to that of the AFM. Interestingly, the MA model performs well in RMSE ($RMSE_{MA} = 0.3633$) and also exhibits a large AUC ($AUC_{MA} = 0.6508$). The high variances again indicate that differences between the AFM, the LG and the MA models are not significant. The DIS and DISMA models perform considerably worse in RMSE and AUC than the best three regression models. The performance of BKT is similar to the first version of the data set, with an RMSE ($RMSE_{BKT} = 0.3613$) in the range of the best regression models and a significantly lower AUC ($AUC_{BKT} = 0.6102$).

3.2 Generalization to new students

In a second experiment, we investigated how well the different regression models generalize to new students using a student-stratified 10-fold cross validation. For new students (i.e., the students in the test set), the number of opportunities to mastery is not known, therefore only the AFM and the LG model were included in this analysis. Prediction accuracy along with standard deviations for the regression models as well as BKT is listed in Tab. 2. The LG model shows the best performance in all error measures for *Version 1* of the data set. The performance of the AFM is very close to that of the LG model in RMSE ($RMSE_{LG} = 0.4164$, $RMSE_{AFM} = 0.4200$). The high variance indicates that there are no significant differences between the two models regarding RMSE. The AUC of the LG model is, however, considerably higher than that of the AFM ($AUC_{LG} = 0.6931$, $AUC_{AFM} = 0.6693$).
Both regression models again outperform BKT in RMSE ($RMSE_{BKT} = 0.4236$) and AUC ($AUC_{BKT} = 0.6688$), but the high variance indicates that there are no significant differences in RMSE between all three models and also not in AUC between the AFM and the BKT model.

Table 2: Prediction accuracy of student-stratified cross-validation for data set *Version 1* and *2*. The values in brackets denote the standard deviations. The best model per error measure is marked (*).

		RMSE	**Accuracy**	**AUC**
Data set: Version 1	AFM	0.4200 (0.0184)	0.7525 (0.0300)	0.6693 (0.0222)
	LG	0.4164 (0.0175)*	0.7583 (0.0248)*	0.6931 (0.0211)*
	BKT	0.4236 (0.0216)	0.7546 (0.0304)	0.6688 (0.0244)
Data set: Version 2	AFM	0.4008 (0.0247)	0.7850 (0.0296)	0.6755 (0.0335)
	LG	0.3936 (0.0241)*	0.7859 (0.0295)*	0.7199 (0.0260)*
	BKT	0.4032 (0.0241)	0.7849 (0.0297)	0.6810 (0.0289)

The results for *Version 2* of the data set show a similar picture. As expected, all models demonstrate a higher prediction accuracy for *Version 2* of the data set. As this version of the data set includes only students that mastered a skill, overall performance is more homogeneous and therefore prediction is easier.

4. DISCUSSION

AFMs are widely used to analyze and improve student learning [5, 15, 23]. However, AFMs are prone to student attrition when applied to data from mastery learning: As students are aligned by opportunity count, the right hand side of the learning curve fitted by an AFM is dominated by students, who require a large number of opportunities to master a skill, which might in turn lead to underestimation of learning rates γ_k. Indeed, [17] observed that averaging over different students with different initial knowledge states and learning rates may result in aggregated learning curves that appear to show little student learning, even though a mastery learning student model such as BKT identified the students as mastering the skills at runtime. This issue can be solved by using alternative models for fitting the learning curves [17]. Our experiments on data from a mastery learning student model (dynamic Bayesian network) with confirmed learning (significant improvement in external post-tests) support these results: AFM fitted positive learning rates γ_k for about half of the skills and only 70% of the positive γ_k were significantly different from zero. Alternative models, such as the LG and MA models predicted positive learning for all skills and learning rates γ_k and generally showed a higher variance, i.e., learning rates differed from skill to skill. Our results demonstrate that different (although very similar) regression models explain the same data in a different way and that alternative regression models predict different patterns of learning.

Despite the different parameter fits, prediction accuracy of the regression models is very similar. When it comes to generalizing to new students, the LG model shows the most accurate prediction. However, as we observe a high variance in accuracy measures, there is most likely no significant difference in prediction accuracy between the AFM and the LG model. Although the AFM performs best in predicting the first re-test, the high variance of the error measures indicates that there is no significant difference between the AFM, the LG and the MA models. The disaggregated models (DIS, DISMA) perform significantly worse than the other regression models. As the disaggregation into different subpopulations increases the number of parameters, the lower performance of these models might be due to overfitting. This hypothesis is supported by the fact that the disaggregated models outperform the other regression models on the training data set in all error measures. Nonetheless, [17] demonstrated the potential of disaggregated models. Prediction accuracy of these models should therefore be evaluated on larger data sets.

BKT models are outperformed by most of the regression models when it comes to prediction accuracy on unseen data. The AFM and the LG model show a higher accuracy when predicting the first re-test, while the AFM, the LG and the MA model generalize better to new students than BKT. Although these differences are probably not significant (due to the high variance in the error measures), they are still interesting. One reason for this observation might be that BKT does not model forgetting. Our analyses have, however, shown that there is forgetting in the data. As the LG and MA models fit only positive learning rates γ_k, this explanation is only valid for the AFM model. Another reason for the superiority of the logistic regression models could be that traditional BKT does not have any student individualization. However, [26] demonstrated on a different data set that a student individualized parameter p_0 does not lead to significant improvements. The reason for the difference in prediction accuracy between BKT and logistic regression models therefore needs to be investigated further.

5. CONCLUSION

In this work, we presented alternative logistic regression models to AFMs, which are able to adjust for mastery-based data sets. Our results demonstrate that the parameter fits for different (although very similar) regression models vary a lot. We also showed that despite the differences in parameter fit, most of the regression models cannot be distinguished regarding prediction accuracy on unseen data. Finally, our evaluations revealed that logistic regression models outperform BKT, when assessing performance in prediction.

In the future, we would like to further analyze the differences between the proposed modeling techniques. Pre-post gain data might be used to evaluate the different logistic regression models. Is the pre-post gain better predicted by improvement on all skills, as per the LG and MA models, or improvement on a subset of skills, as per the AFM. It could be that, as with the retention measure, AFM somewhat under predicts learning gain and LG somewhat over predicts learning gain. Furthermore, we would like to analyze the differences in prediction between BKT and logistic regression models.

6. ACKNOWLEDGMENTS

This work was funded by the CTI-grant 11006.1.

7. REFERENCES

[1] 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*, pages 52–63, 2010.

[2] P. Boeck. Random Item IRT Models. *Psychometrika*, 73(4):533–559, 2008.

[3] N. E. Breslow and D. G. Clayton. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

[4] E. Brunskill. Estimating Prerequisite Structure From Noisy Data. In *Proc. EDM*, pages 217–222, 2011.

[5] 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*, pages 511–518, 2007.

[6] H. Cen, K. R. Koedinger, and B. Junker. Comparing Two IRT Models for Conjunctive Skills. In *Proc. ITS*, pages 796–798, 2008.

[7] K.-M. Chang, J. Beck, J. Mostow, and A. Corbett. A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems. In *Proc. ITS*, pages 104–113, 2006.

[8] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *UMUAI*, 4(4):253–278, 1994.

[9] J. P. González-Brenes and J. Mostow. Dynamic Cognitive Tracing: Towards Unified Discovery of Student and Cognitive Models. In *Proc. EDM*, pages 49–56, 2012.

[10] J. P. González-Brenes and J. Mostow. Topical Hidden Markov Models for Skill Discovery in Tutorial Data. *NIPS - Workshop on Personalizing Education With Machine Learning*, 2012.

[11] T. Käser, G.-M. Baschera, J. Kohn, K. Kucian, V. Richtmann, U. Grond, M. Gross, and M. von Aster. Design and evaluation of the computer-based training program Calcularis for enhancing numerical cognition. *Front. Psychol.*, 2013.

[12] T. Käser, A. G. Busetto, G.-M. Baschera, J. Kohn, K. Kucian, M. von Aster, and M. Gross. Modelling and optimizing the process of learning mathematics. In *Proc. ITS*, pages 389–398, 2012.

[13] T. Käser, A. G. Busetto, B. Solenthaler, G.-M. Baschera, J. Kohn, K. Kucian, M. von Aster, and M. Gross. Modelling and Optimizing Mathematics Learning in Children. *IJAIED*, 23(1-4):115–135, 2013.

[14] K. Koedinger and E. McLaughlin. Seeing language learning inside the math: Cognitive analysis yields transfer. In *Proc. of the 32nd Annual Conference of the Cognitive Science Society*, pages 471–476, 2010.

[15] K. Koedinger, J. Stamper, E. McLaughlin, and T. Nixon. Using Data-Driven Discovery of Better Student Models to Improve Student Learning. In *Proc. AIED*, pages 421–430, 2013.

[16] J. I. Lee and E. Brunskill. The Impact on Individualizing Student Models on Necessary Practice Opportunities. In *Proc. EDM*, pages 118–125, 2012.

[17] R. Murray, S. Ritter, T. Nixon, R. Schwiebert, R. Hausmann, B. Towle, S. Fancsali, and A. Vuong. Revealing the Learning in Learning Curves. In *Proc. AIED*, pages 473–482, 2013.

[18] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *Computer Journal*, 7:308–313, 1965.

[19] Z. A. Pardos and N. T. Heffernan. Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. In *Proc. UMAP*, pages 255–266, 2010.

[20] Z. A. Pardos, S. Trivedi, N. T. Heffernan, and G. N. Sárközy. Clustered knowledge tracing. In *Proc. ITS*, pages 405–410, 2012.

[21] P. I. Pavlik, H. Cen, and K. R. Koedinger. Performance Factors Analysis - A New Alternative to Knowledge Tracing. In *Proc. AIED*, pages 531–538, 2009.

[22] J. Reye. Student Modelling Based on Belief Networks. *IJAIED*, 14(1):63–96, 2004.

[23] J. C. Stamper and K. R. Koedinger. Human-machine Student Model Discovery and Improvement Using DataShop. In *Proc. AIED*, pages 353–360, 2011.

[24] Y. Wang and N. T. Heffernan. The student skill model. In *Proc. ITS*, pages 399–404, 2012.

[25] M. Wilson and P. De Boeck. Descriptive and explanatory item response models. In P. De Boeck and M. Wilson, editors, *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag, 2004.

[26] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon. Individualized Bayesian Knowledge Tracing Models. In *Proc. AIED*, pages 171–180, 2013.