Items, skills, and transfer models: which really matters for student modeling?

Y. GONG AND J. E. BECK
Worcester Polytechnic Institute, U.S.A.

Student modeling is broadly used in educational data mining and intelligent tutoring systems for making scientific discoveries and for guiding instruction. For both of these goals, having high model accuracy is important, and researchers have incorporated a variety of features into student models. However, since different techniques use various features, when evaluating those approaches, we could not easily figure out what is key for a high predictive accuracy: the model or the features. In this paper, to establish such knowledge, we performed empirical studies varying which features the models considered such as items, skills, and transfer models. We found that item difficulty is a better predictor than skill difficulty or student proficiencies on the transfer model. Moreover, we evaluated two versions of the PFA model; the one with item difficulty resulted in slightly higher predictive accuracy than the one with skill difficulty. In addition, prior work has shown that considering student overall proficiencies, not just those thought to be important by the transfer model, works substantially better on ASSISTments data. However, in this study, we failed to find consistency of this phenomenon on the data collected from the Cognitive Tutor.

Key Words and Phrases: Performance factors analysis, item difficulty, student performance, predictive accuracy

1. INTRODUCTION

Student modeling has been broadly used in educational data mining and applications of intelligent tutoring systems (ITS) for discovering scientific truth about student knowledge, performance, behaviors and motivations, with the goal of leading to a better understanding of students. A wide array of research has been conducted based on student modeling, such as research related to "Gaming the system" [2, 10], the impacts of student non-academic strengths on learning [1, 11], and the effect of item order on student learning [16]. Furthermore, a good student model is also indispensable for a successful ITS. Given the effectiveness of ITS [9, 15] , findings such as one-to-one tutoring is better than classroom tutoring [3], and that a step-based computer tutor was not outperformed by human tutors [7], give us a sense that a reason for an ITS's success is its ability to provide individualized tutoring (one-to-one tutoring). Such tutoring relies on the support of an accurate student model in order to understand students.

Our research interest in this paper lies in student modeling. We simply wish to study what makes a good student model. There is more than one criterion for judging the goodness of a student model [21]. In this study, we focus on the student model's predictive accuracy. Although student models are frequently evaluated, it can be difficult to know what aspect is responsible for a success or failure. As a result, knowledge as to what makes an accurate student model is insufficient. Our goal in this study is to use the same student modeling framework for different evaluations, to construct guidance about what features (student model components) are important for designing an accurate student model.

There are many potential features that can inform a student model. In this study, items, skills and transfer models were chosen for evaluation, as those are the most commonly used components across different student modeling techniques. In addition, it is also meaningful to examine complete student models constructed with those features, as knowledge about whether and how much multiple features can contribute higher accuracy is also significant. Therefore, we evaluated a series of student models.

1.1. STUDENT MODELING FRAMEWORK

Authors' addresses: {Y. Gong, J. E. Beck}, Computer Science Department, Worcester Polytechnic Institute, U.S.A. E-mail: {ygong, josephbeck}@wpi.edu.

Performance Factors Analysis (PFA) is a student modeling approach proposed by Pavlik, et al. in 2009 [19]. It takes the form of logistic regression with student performance as the dependent variable. We chose PFA as our framework as, relative to Bayesian networks, logistic regression is more flexible to incorporate more (or different) predictors.

It is particularly important to note that there are two student models, both of which were named as Performance Factors Analysis. Both models were designed based on the reconfigurations of Learning Factors Analysis [4] by dropping student variable and considering a student's prior correct and incorrect performances. The two models vary in their independent variables. The model presented in [20] estimates item difficulty (i.e. one parameter per question); the other [19] estimates skill difficulty (i.e. one parameter per skill. Note that in the original paper [19], the authors used the term "knowledge components (KC)" while we use the term "skills"). In this paper, we refer to the first model as the PFA-item model; the other is represented as the PFA-skill model.

$$m(i, j \in required\_skills, q \in questions, s, f) = \beta_q + \sum_{j \in required\_skills} (\gamma_j s_{i,j} + \rho_j f_{i,j}) \quad \text{(1) PFA-item}$$

$$m(i, j \in required\_skills, s, f) = \sum_{j \in required\_skills} (\beta_j + \gamma_j s_{i,j} + \rho_j f_{i,j}) \quad \text{(2) PFA-skill}$$

The *m*s in Equation 1 and 2 are logits (i.e., are transformed by $e^x/(1+e^x)$ to generate a probability). They represent the likelihood of student *i* generating a correct response to an item. In the equations, $s_{i,j}$ and $f_{i,j}$ are two observed variables, representing the numbers of the prior **s**uccessful and **f**ailed practices done by student *i* on skill *j* . The corresponding two coefficients (γ_j and ρ_j) are estimated to reflect the effects of a prior correct response and a prior incorrect response of skill *j*. Rather than considering all of the skills in the domain, the PFA model focuses on just those skills required to solve the problem.

The PFA-item model estimates a parameter (β_q) for each question representing its difficulty. In the PFA-skill model, as seen in Equation 2, the β parameter has a subscript of *j*, indicating that it captures the difficulty of a skill. Also, it is moved to the inside of the summation part to incorporate multiple skills, i.e., in PFA-skill an item's difficulty is the sum of its skills' difficulties.

1.2. EXPERIMENTS

The data used in this study are a small portion of the algebra-2005-2006 development data set for the KDD cup competition 2010 from the Cognitive Algebra Tutor. Since the original data set is very large, to form our working data set, we randomly selected 74 students and their performance records, 94,585 steps completed by the students. We don't have access to the transfer model used in this data set. Thus for determining which skills are required in a question, we directly used the skill labels given in the data. There are a number of questions that do not specify which skills are required to solve them. For those questions, we removed them from the data set. Therefore, in the remaining data set, there are 117 algebra skills, including: Addition/Subtraction, Remove constant, Using simple numbers, Using small numbers, etc.

In this study, we did 4-fold crossvalidation at the level of students, and tested the models on held-out students. We chose to hold out at the student level since that results in a more independent test set. We focused on a student model's accuracy in predicting those held-out students' performances. Predictive accuracy is the measure of how well the instantiated model fits the test data. We used two metrics to examine the model's predictive performance on the test data set: Efron's R^2 and AUC of ROC curve (Area under the curve of Receiver Operating Characteristic). Efron's R^2 is a measure of how

much error the model makes in predicting each data point, compared to a model that uses the mean of the those data to predict. A 0 indicates the model does no better than simply predicting the mean; a 1 indicates prefect prediction. A negative value of Efron's R^2 indicates that the model has more error than a model that just simply guesses the mean for every prediction. AUC of the ROC curve evaluates the model's performance on classifying the target variable which has two categories. In our case, it measures the model's ability to differentiate students' positive and negative responses. AUC of 0.5 is the baseline, which indicates random prediction.

In the result section, we report the comparative results by providing the R^2 and AUC measurements across all four folds. To test the differences of the means, we also performed paired two-tailed t tests using the results from the crossvalidation with degrees of freedom of N-1, where N is the number of folds (i.e. df=3).

2. STUDENT MODEL COMPONENTS

Many student model components could be important for enabling a student model to achieve high accuracy in predicting student performance.

Student proficiencies on required skills are widely used in many student modeling techniques [4, 5, 19, 20]. Since the transfer model is responsible for providing which skills are required to solve a problem, we refer to "using student proficiencies on required skills to predict" as "using transfer models to predict". The transfer model is often treated as the primary component in student modeling, so is the first component we considered.

Our question was simple: how much variance do transfer models account for? Specifically, how much can a model's predictive accuracy benefit from observing a student's prior performances on required skills? To answer this question, we designed a model that solely considers student proficiencies on the transfer model. We accomplished the model on the basis of the PFA-item model by removing the predictor, item difficulty (β_q), from Equation 1, for the reason that item difficulty is not related to the transfer model. Therefore, the new model has student performances on a series of question as the single predictor, so the only variable predicting the possibility of a student's correct is his proficiencies on required skills.

Item difficulty (question difficulty) has been less studied in student modeling, but is used in Item Response Theory (IRT) [22], a generally effective technique for assessing students [8, 22] such as for computer-based testing [6, 14]. Therefore, it is reasonable to infer that item difficulty is an important predictor of student performance. Item difficulty hasn't been widely used in student modeling until recently when the PFA-item model was proposed [20], as well being integrated into Knowledge Tracing [5] in order to better predict student performance [18]. Hence, in student modeling, there were few attempts for exploring the ability of item difficulty to accurately predict student performance.

Similar to how we test the effect of the transfer model in isolation, in order to test the effect of item difficulty we modify the PFA-item model by dropping the part corresponding to student proficiencies (the part inside the Σ in Equation 1). So the model only has the parameter β_q. Since the model has excluded other features, it can be used to discover the pure ability of item difficulty to contribute the model's predictive accuracy.

The last component we are interested to see is skill, rather than item, difficulty. It is also not commonly used, although Learning Factors Analysis [4] uses skill difficulty in the model. Since the PFA-skill model was reconfigured based on the LFA model, it inherits this feature. To examine skill difficulty, we built a model based on the PFA-skill model (Equation 2) and removed the part corresponding to student proficiencies. Only the skill difficulty parameter (β_j) after the sigma sign is left to capture the effect of the required skills for the question.

2.1. RESULTS

In this section, we examine the predictive power provided by different student model components, including item difficulty, skill difficulty and student proficiencies on the skills in the transfer model. Since each of our models only consider a single feature, the results of testing the model can be attributed to that component.

With respect to modeling item difficulty, we were forced to make a compromise when designing the models. Due to a characteristic of the Cognitive Tutor data, it is not sensible to use the question's identity. In the Cognitive Tutor, a question can have multiple steps, each of which typically requires different skills. Therefore, in the Cognitive Tutor, if a question identity occurs multiple times in the student performance records, we cannot simply assume that they concern the same question. For example, a record might be the first step of a question, while another record with the same question identity might be the tenth step of the question. The difficulties of the two steps are probably not the same as they involve different skills and different aspects of the question. For modeling skill difficulty, there is no difficulty, but it presents clear problems for modeling item difficulty. A solution is to build a new question identity combining the original question identity and the skills required in a step [18]. For instance, if the original question id is Q1 and the first step of the question requires "Addition", we can build a new question id, Q1-Addition; while if the tenth step requires "Using small numbers", we have another question id, Q1-UsingSmallNumbers. However, this way results in a very large number of question identities, over 8000 in our data, and it causes a severe computational problem for logistic regression and an inability to fit the model within SPSS, even with increased memory. Therefore, we made a pragmatic decision: for each step, we represented its difficulty using the summation of the difficulty of the original question and the difficulties of the required skills in that step. In this way, the computational cost is greatly reduced and an approximate difficulty for the step can be estimated. The corresponding equation is shown Equation 3.

$$m(i,j,q \in questions,s,f) = \beta_q + \sum_{j \in required\_skills} (\beta_j + \gamma_j s_{i,j} + \rho_j f_{i,j}) \quad (3)$$

The computationally viable method

Table I shows the comparative results of models, each of which was fit by a single student model component. First, we found that compared to the other student model components, the model using item difficulty results in higher predictive accuracy and the differences in the means are significant. In the comparison of item difficulty vs. skill difficulty, the t-tests resulted in $p=0.02$ in R^2 and $p=0.005$ in AUC. In the comparison between the model using item difficulty and the model using transfer models, the t-tests yielded $p=0.006$ in R^2 and $p=0.48$ in AUC. The p-value in AUC suggests that there is not enough evidence to show that the two models have different classification abilities for the student performances, while the predictive error made by the model using item difficulty is significantly smaller than its counterpart.

Table I Comparative performance on unseen students

Student model component	**R^2**	**AUC**
Item difficulty	0.149	0.739
Skill difficulty	0.139	0.720
Student proficiencies on the transfer model	0.132	0.738

The results concerning item difficulty suggest that contrary to the traditional belief that student proficiencies on the transfer model (required skills) are the most important

predictor, instead item difficulty is an even more powerful predictor of student performance. This finding is also consistent with the finding in the study using the data gathered from ASSISTments [13], suggesting that item difficulty can cover more variance of student performance is a general phenomenon across different computer tutors and different populations.

Table I also shows the results of comparing skill difficulty and student proficiencies on the transfer model. The results of the two metrics do not agree with each other, but both differences are found to be reliable: $p=0.03$ in R^2 and $p=0.02$ in AUC; therefore, it is still uncertain about whether skill difficulty or student proficiency is more important for predicting student performance.

3. STUDENT MODELS

Aside from getting knowledge about how components perform in isolation, it is also important to understand the predictive accuracy of complete models using multiple features, such as the full PFA-item model (Equation 1). It makes sense to examine a complete model as a whole for the following two reasons. First, from a scientific point of view, it is interesting to find out whether different features account for unique variation in predicting student behavior, or whether one feature largely subsumes another. Second, from a practical point of view, knowing whether adding a certain feature is a positive step for improving the model's predictive accuracy helps design a compact, yet effective student model.

3.1. THE TWO VERSIONS OF THE PFA MODEL

We examine the two PFA models, PFA-item and PFA-skill, because direct comparisons between these two have never been performed.

When the PFA-skill model was presented, the designers of the model, using data from Cognitive Tutors, performed evaluations against a well-established student model, Knowledge Tracing, and found that on the student population of Cognitive Tutor, the PFA-skill model is somewhat superior to KT [19]. On the other hand, our prior work applied the PFA-item model to another tutor, ASSISTments, and found that the PFA-item model was markedly superior to KT [12]. Since there have been no studies comparing PFA-item and PFA-skill at the same time and on the same population, we are unsure about the reason for this difference of results.

3.2. A VARIANT OF THE PFA MODEL: THE OVERALL PROFICIENCIES MODEL

We proposed the *overall proficiencies* model, a variant of the PFA-item model, in prior work [13]. This model incorporates the idea that student proficiencies on all skills, not just those the transfer model thinks are required for a particular item, could be important for better predicting student performance on the item. Prior work found that this model performed significantly better than the PFA-item model on ASSISTments data [13]. In this study, we wanted to extend this model to another tutoring environment, Cognitive Tutor, and another population, students of Cognitive Tutor. Since there are many differences between the two systems, we aimed to use this study to better understand the overall proficiencies model.

We had two hypotheses to support the reasonableness of the overall proficiencies model. The first is that the assumption of using transfer models to predict might not always hold, as transfer models assume that only student proficiencies on the required skills have impact on question solving. In other words, student proficiencies on non-required skills are independent of student performance on the problem. However, it is not always true for all ITSs, perhaps due to the possibility that there are relationships between required skills and non-required skills, which are not well captured by the

transfer model; or perhaps problems involve a broader range of skills than the subject matter expert believed and encoded in the transfer model. Second, since in some student modeling techniques, student ability is viewed as a factor helpful for producing higher model accuracy [4, 17], we assume that a student's overall proficiencies can be treated as a sign to reflect the student's overall ability. Thus using those is able to provide the model more information about the student, so as to enable the model to reach higher predictive accuracy.

The overall proficiencies model is built based on the PFA-item model. We modified the PFA-item model's predictors by replacing *REQUIRED_skills* with *ALL_skills* the subscript on the Σ). The equation is shown as follows.

$$m(i, j \in ALL\_skills, q \in questions, s, f) = \beta_q + \sum_{j \in ALL\_skills} (\gamma_j s_{i,j} + \rho_j f_{i,j}) \quad (4)$$

The overall proficiencies model

3.3. RESULTS

In this section, we compare student models which consider a selected set of student model components. Table II shows the mean performance and per-fold performance for each model and metric. Note that both PFA-item and PFA-skill both outperform the item difficulty model in Table I. Since the transfer model is needed to train both of the PFA models (to get the success and failure counts on each required skill), there is evidence that transfer models are in fact helpful for student modeling.

Table II Model performance on test data

	PFA-item		**PFA-skill**		**Overall proficiencies**	
	R^2	**AUC**	**R^2**	**AUC**	**R^2**	**AUC**
Fold 1	0.194	0.768	0.181	0.756	0.090	0.694
Fold 2	0.177	0.762	0.179	0.760	0.035	0.709
Fold 3	0.144	0.756	0.142	0.748	-0.178	0.674
Fold 4	0.149	0.746	0.143	0.740	-0.082	0.660
Mean	**0.166**	**0.758**	**0.161**	**0.751**	**-0.034**	**0.684**

The first comparison is between the PFA-item model and the PFA-skill model. We noticed that both models' predictive accuracies vary considerably across 4 folds. This similar trend is shown in both R^2 and AUC. In the PFA-item model, the R^2 values vary from as large as 19.4% to as small as 14.4% (standard deviation of 0.024 across folds). Prior study has applied the PFA-item model to ASSISTments data, but found less variation (standard deviation of 0.008 across folds). This finding suggests that across students, student performances of this data set of Cognitive Tutor have larger variance than that of ASSISTments, which is possibly because that there were only 18 or 19 students in each fold, and thus potentially making the model's performance unstable.

Second, between the PFA-item and the PFA-skill models, the means of the two measurements suggest that the PFA-item model seems to outperform the PFA-skill model, but the p-value of R^2 is 0.22, while that of AUC is 0.051. The p value of 0.22 indicates that when comparing the two models in terms of their abilities to minimize error during predictions, we were not able to reject the null hypothesis that the two models achieved different predictive accuracy. In the classification ability, the PFA-item model is marginally reliably better than the PFA-skill model, suggested by p=0.051 in AUC.

Third, it is worth pointing out that on this data set, the PFA-item model produces many more parameters than the PFA-skill model. Since we used a compromised approach to implement the PFA-item model, there are around 950 more parameters (each

per original question identity). If we implemented the model in its original way, it would have around 8000+ parameters (each per created question identity). As a consequence of having additional parameters, the PFA-item model is prone to overfitting. To demonstrate overfitting, in Table III we report the R^2 on the *training* data for each fold. For each model, we compared the R^2 values on training data with the R^2 values on test data. We found that compared to the PFA-skill model, the PFA-item model's performance dropped considerably. Given that the two models performed closely on the test data, the better performance on training data of the PFA-item model did not transfer to test data, suggesting overfitting occurred. However, perhaps with a larger dataset the models' training- and test-set performances would be more similar.

Table III Model performance (R^2) on training data

	PFA-item	**PFA-skill**	**Overall proficiencies**
Fold 1	0.229	0.185	0.234
Fold 2	0.284	0.241	0.286
Fold 3	0.231	0.187	0.232
Fold 4	0.237	0.192	0.240
Mean	0.245	0.201	0.248

In this study, we also applied the overall proficiencies model on the Cognitive Tutor data. Interestingly, the model did best in all four folds on the training data, shown in the last column of Table III, but performed the worst on the test data, shown in the last two columns of Table II. Furthermore, the results in Table II have more variability than the PFA-item and PFA-skill models, indicating that the overall proficiencies model performed even more unstably on different students. The results suggest that the overall proficiencies model on the Cognitive Tutor data has serious overfitting problems, and is not suitable for their student records, at least with amount of data used in this study. More discussions about the potential reasons are presented in the section of future work.

4. CONTRIBUTIONS

This study performed explorations of student modeling and contributed basic knowledge to the community.

First, we provided insights in terms of what student model components matter for building an accurate student model of student performance. Different student model components have been used in various student modeling techniques [4, 5, 13, 19, 20], yet thorough inspections of the effectiveness of those components on producing accurate predictions were missing. As a replication and extension of our prior work [13], this work considered one more student model component, skill difficulty, and also tested student model components on another population: students of Cognitive Tutor. Similar to our previous finding, item difficulty is more accurate for predicting student performance than student proficiencies on skills related to the problem. The finding is important, especially given that student proficiencies on related skills are widely used in almost all well-established student modeling techniques. However, using item difficulty can result in a painful model fitting process, depending on the number of items in the data set. Take PFA as example, logistic regression is particularly time-consuming in the presence of a large number of predictors. Therefore, we suggest that although item difficulty works better for forming an accurate student model, decisions should be made based on concrete characteristics of the data, especially given that, for the Cognitive Tutor data, item difficulty only slightly outperformed skill difficulty.

Second, Performance Factors Analysis refers to two different models. In this paper we differentiated them as the PFA-item model and the PFA-skill model. The PFA-skill

model was evaluated against KT and found to be somewhat better [19]; while the PFA-item model was compared with KT as well, but shown with substantially better performance [12]. The direct comparison between the two models has never been performed, leading to uncertainty about their relative performance. In this study, we found that on the Cognitive Tutor data, the PFA-skill model is slightly worse than the PFA-item model, yet with much fewer parameters estimated. The PFA-item model by contrast, for our data set, estimates a large number of parameters. Even with the restricted to be computationally tractable method, it still produced 900+ more parameters, which resulted in a relative 3% improvement. In addition, the PFA-item model is more prone to overfitting. Our results suggest that the PFA-skill model is a good option for predicting student performance data similar to the Cognitive Tutor data.

Finally, we proposed a variant of the PFA model, the overall proficiencies model, in our prior work and showed that the model works substantially better than PFA-item on ASSISTments data [13]. Therefore, applying the model to data from a different tutor environment and a different student population helps achieve a deeper understanding of this new model. We found that the similar trend was not observed on Cognitive Tutor data, as the overall proficiencies model performed poorly on the test data, indicating that the model cannot be generalized on those held-out students. The results suggest the overall proficiencies model does not universally result in a stronger model fit. We have a number of hypotheses for what characteristics would be, so the detailed conditions that make the model perform better are still uncertain for us.

5. FUTURE WORK AND CONCLUSIONS

This study creates several unanswered questions that motivate further research work.

To establish the fundamental knowledge with respect to what component matters for a student model, broader inspections of the components involving different experimental populations and different tutors are still needed, especially given the uncertainty of whether skill difficulty and student proficiencies on the transfer model is able to produce more accurate prediction. In addition, since ASSISTments has several different features from Cognitive Tutor in its pedagogical policies, transfer models, student population, etc., it is meaningful to test the PFA-skill model on the ASSISTments data to see whether it is comparable to the PFA-item model, or whether the differences between the tutors cause one model to outperform the other.

We have no clear answers to explain what major differences between the Cognitive Tutor data and the ASSISTments data cause so different predictive performances of the overall proficiencies model. As we hypothesized in prior study [13], there were at least two potential reasons for the success of the model.

First, the transfer model used in ASSISTments might not be specific enough to explicitly designate all associations between a question and its required skills. Thus, student proficiencies on non-required skills are not independent of the proficiencies on required ones. In other words, there might be relationships between required and non-required skills. Given that the model performed poorly on the Cognitive Tutor data, we think it is due to the following two conditions of the Algebra Cognitive Tutor.

1. The comprehensiveness and correctness of the transfer model.

In fact, the domain expert of ASSISTments intensively encoded a smaller range of skills in the transfer model, assumed that the prerequisite skills are required by default, and thus did not indicate them in the transfer model. Therefore, in ASSISTments, if a question requires Pythagorean Theorem, it is highly likely that it also requires equation solving and square root, but the relationships are not captured by the transfer model. The Cognitive Tutor by contrast, has much more meticulous representation. For example, it

has skills such as “Remove constant” in equation solving, “Remove coefficient” in equation solving, “Entering a given”, etc. Those skills are all hidden beneath a single skill “equation solving” in ASSISTments. Specifically, there are 104 mathematical skills in ASSISTments, covering five strands of middle school Math: algebra, geometry, measurement, number sense and data analysis. By contrast, the Cognitive Tutor has 110 skills just for algebra. The comprehensive transfer model of Cognitive Tutor might be a reason to cause the overall proficiencies model to lose its advantage to deal with implicitly existing relationships between required and non-required skills. An additional factor is the degree of knowledge engineering. The Cognitive Tutors' transfer models have been refined over years of experiments, while ASSISTments transfer models were made similarly to most ITS: a subject-matter expert designed them. Although we lack data, we suspect the Cognitive Tutor's transfer models are more accurate, and this factor could certainly impact which student modeling approach works better.

2. The way of tutoring

In ASSISTments, a student enters a single answer to an item, and only has to answer subsidiary “scaffolding” questions in the event the student answers a main question incorrectly. In contrast, in the Cognitive Tutor, no scaffolding questions (steps) are allowed to be skipped. A main question in ASSISTments typically asks higher abstract-level skills, i.e. ask all detailed skills at once; while its scaffolding questions test more specific skills. Thus, flexibly accessing to scaffolding questions causes the model to miss chances to observe student performance associated with fine-grained skills. Consider that if a student makes a successful practice on a skill, it is likely that the student's knowledge on many other skills benefits from it as well, and just simply we don't have the chance to observe that. Contrariwise, Cognitive Tutor forces a question to be broken down into steps, so it is not possible for the model to miss any observations of a student practicing on any skills; a correct response of a skill probably has little impact on other skills.

Second, since scaffolding questions are not always used, there were fewer observations of students solving problems that test individual skills in the ASSISTments tutor [13]. Therefore, the student overall proficiencies provides useful evidence to the model to enable the model to more accurately predict. For the Cognitive Tutor data used in this study, due to solving each step being mandatory, there were many more observations for each skill. In addition, within the Cognitive Tutor there was more intensive usage by students. Specifically, for fine-grained algebra skills of the Cognitive Tutor, there were approximately ~100 observations per student per skill; in ASSISTments, with its 104 coarser-grained skills, there were on average fewer than 10 observations per student per skill. Therefore, for the Cognitive Tutor data, dense evidence for a student's performance on those fine-grained skills might also be a reason for the poor performance of the overall proficiencies.

In summary, this study explored what matters for a student model in terms of producing higher accuracy in predicting student performance. Consistent with our prior finding, for predictive accuracy. item difficulty outperformers transfer models, the most widely used student model components, as well as skill difficulty. The comparisons between the PFA-item and the PFA-skill models brought up an insight that the PFA-skill model is slightly worse than the PFA-item model, but has fewer parameters, a smaller problem of overfitting, and is much more computationally tractable. We extended the overall proficiencies model to the data collected from Cognitive Tutor and found it performed worse than the PFA-item model, suggesting that the overall proficiencies model works well only under certain conditions of an ITS, an area that needs additional exploration.

ACKNOWLEDGEMENTS

For the full list of over a dozen funders please see http://www.webcitation.org/5xp605MwY.

REFERENCES

[1] ARROYO, I., and WOOLF, B. 2005. Inferring Learning and Attitudes from a Bayesian Network of Log File Data. Proceedings of the 12th International Conference on Artificial Intelligence in Education. pp.33-40.

[2] Baker, R.S., Corbett, A.T. and Koedinger, K.R. (2004) Detecting Student Misuse of Intelligent Tutoring Systems. Proceedings of the 7th International Conference on Intelligent Tutoring Systems, 531-540.

[3] Bloom, B.S.: The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher 13, 4-16 (1984)

[4] Cen, H., Koedinger, K. and Junker, B.: Learning Factors Analysis - A General Method for Cognitive Model Evaluation and Improvement. Proceedings of the 8th International Conference on Intelligent Tutoring Systems. pp. 164-175. (2006)

[5] Corbett, A. & Anderson, J. (1995) Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction. 4: pp. 253-278.

[6] www.ets.org

[7] Evens, M., Michael, J.: One-on-one Tutoring By Humans and Machines. Erlbaum, Mahwah (2006)

[8] Feng, M., Heffernan, N.T., & Koedinger, K.R. (2009). Addressing the assessment challenge in an Intelligent Tutoring System that tutors as it assesses. The Journal of User Modeling and User-Adapted Interaction. Vol 19: p243-266.

[9] Feng, M., Heffernan, N. & Beck, J.(2009) Using Learning Decomposition to Analyze Instructional Effectiveness in the ASSISTment System. Proceedings of the 2009 Artificial Intelligence in Education Conference. IOS Press. pp. 523-530.

[10] Gong, Y., Beck, J., Heffernan, N. T.& Forbes-Summers, E. (2010)The impact of gaming (?) on learning at the fine-grained level. In Aleven, V., Kay, J & Mostow, J. (Eds) Proceedings of the 10th International Conference on Intelligent Tutoring Systems (ITS2010) Part 1. Springer. Pages 194-203.

[11] Gong, Y., Rai, D. Beck, J. E. & Heffernan, N. T. (2009) Does Self-Discipline impact students' knowledge and learning? In Barnes, Desmarais, Romero & Ventura (Eds) Proc. of the 2ndInternational Conference on Educational Data Mining. Pp. 61-70. ISBN: 978-84-613-2308-1.

[12] Gong, Y, Beck, J. E., Heffernan, N. T. (2010) How to Construct More Accurate Student Models: Comparing and Optimizing Knowledge Tracing and Performance Factors Analysis. International Journal of Artificial Intelligence in Education. Accepted, 2010.

[13] Gong, Y., Beck,J. E. (Accepted) Looking beyond transfer models: finding other sources of power for student models,. Accepted to the 19th International Conference on User Modeling, Adaptation and Personalization. Girona, Spain.

[14] www.grockit.com

[15] Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. International Journal of Artificial Intelligence in Education, 8, 30-43.

[16] Pardos, Z.A., Heffernan, N.T. (2009). Determining the Significance of Item Order In Randomized Problem Sets. In Barnes, Desmarais, Romero & Ventura (Eds) Proc. of the 2nd International Conference on Educational Data Mining. pp. 111-120. ISBN: 978-84-613-2308-1

[17] Pardos, Z. A., Heffernan, N. T. Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. The 18th Proceedings of the International Conference on User Modeling, Adaptation and Personalization. (2010)

[18] Pardos, Z. A., Heffernan, N. T. (Accepted) KT-IDEM: Introducing Item Difficulty to the Knowledge Tracing Model. Accepted to the 19th International Conference on User Modeling, Adaptation and Personalization. Girona, Spain.

[19] Pavlik, P. I., Cen, H. & Koedinger, K. (2009) Performance Factors Analysis - A New Alternative to Knowledge. Proceedings of the 14th International Conference on Artificial Intelligence in Education, Brighton, UK, pp. 531-538.

[20] Pavlik, P. I., Cen, H., Koedinger, K. : Learning Factors Transfer Analysis: Using Learning Curve Analysis to Automatically Generate Domain Models. Proceedings of the 2rd International Conference on Educational Data Mining. pp.121-130. (2009)

[21] Rai, D, Gong, Y, Beck, J. E. : Using Dirichlet priors to improve model parameter plausibility. Proceedings of the 2nd International Conference on Educational Data Mining, Cordoba, Spain, pp141-148.2009

[22] van der Linden, & W. J., & Hambleton, R. K. (eds.) (1997). Handbook of modern item response theory. New York, NY: Springer Verlag.