

# Analyzing Fine-Grained Skill Models Using Bayesian and Mixed Effects Methods

Zachary A. Pardos, Mingyu Feng, Neil T. Heffernan, Cristina Linquist-Heffernan  
*Worcester Polytechnic Institute*  
{zpardos, mfeng, nth}@wpi.edu

**Abstract.** Two modeling methods were employed to answer the same research question of how accurate the various grained models with 1, 5, 39 and 106 skills are at assessing student knowledge in the ASSISTment online tutoring system and predicting their performance on the 2005 state MCAS test. One method, used by the second author, is mixed-effects statistical modeling. The first author evaluated the problem with a Bayesian networks machine learning approach. We compare the two results to identify benefits and drawbacks of either method and to find out if the two results agree. We report that both methods showed compelling similarity in results especially with regard to residuals on the test. Our analysis of these residuals and our online skills allows us to better understand our model and conclude with recommendations for improving the tutoring system, as well as implications for state testing programs.

## 1. Introduction

Intelligent Tutoring Systems (ITS) rely on models that associate the various skills students are learning with different questions or actions. We created 3 different models with different grain sizes. One model has 5 skills, another 39 and our finest grain model has 106 skills. A model with a single skill was used to represent unidimensional assessment. We found that working with the 20 teachers that use our system, many appreciate the reports made possible by the fine grain sized models that tell them which specific skills a student is doing poorly on. But are these finer grained models at least as accurate as more traditional, course grained models [6] in prediction performance?

To our knowledge, no one else has specifically investigated this question. However some have investigated the results of skill hierarchies using simulated users [2, 3]. This paper attempts to also compare two different ways of modeling skills; Bayesian networks [10] popular in Artificial Intelligence research and mixed-effects modeling [12] popular in statistical departments. We investigate if both modeling methodologies yield similar results to the question of “Are finer grain sized skill models more accurate at test prediction.” We will be able to gain confidence in both types of modeling if one method corroborates the other's results.

## 2. The Massachusetts Comprehensive Assessment System (MCAS)

The MCAS is a Massachusetts state administered standardized test that produces tests for English, math, science and social studies for grades 3rd through 10th. We are focused on only 8th grade mathematics. Our work relates to the MCAS in two ways. First we have built our content based upon the ~300 publicly released items from previous MCAS math tests. Secondly, we will be evaluating our models by predicting the 8th grade 2005 MCAS test which was taken by students after the online data being used was collected.

## 3. Background on the ASSISTment Project

The ASSISTment system is an e-learning and e-assessing system [5]. In the 2004-2005 school year, 600+ students used the system about once every two weeks. Eight math teachers from two schools would bring their students to the computer lab, at which time students would be presented with randomly selected MCAS test items. Each tutoring item, which we call an ASSISTment, is based upon a publicly released MCAS item which we have added “tutoring”, also known as “scaffolding”, to. If students get the item correct they are advanced to the next question. If they answer incorrectly, they are provided with a small “tutoring” session where they are asked to answer a few questions that break the problem down into steps. The first scaffolding question appears only if the student gets the item wrong. We believe that the ASSISTment system has a better chance of showing the utility of fine-grained skill modeling due to the fact that we can ask scaffolding questions that break the problem down in to parts and attempt to identify which skills were to blame. Most MCAS questions that were presented as multiple-choice were converted into text-input questions to reduce the chance of guess. As a matter of logging, the student is only marked as getting the item correct if they answer the question correctly on the first attempt and do not ask for hints.

#### 4. Creation of the Fine-Grained Skill Model

In April of 2005, we staged a 7 hour long “coding session” at Worcester Polytechnic Institute (WPI), where our subject-matter expert, Lindquist-Heffernan, with the assistance of the 3rd author, set out to make up skills and tag all of the existing 8th grade MCAS items with these skills. There were about 300 released test items for us to code. Because we wanted to be able to track learning between items, we wanted to come up with a number of skills that were somewhat fine-grained but not too fine-grained such that each item had a different skill. We therefore imposed upon our subject-matter expert that no one item would be tagged with more than 3 skills. She gave the skills names, but the real essence of a skill is what items it was tagged to. To create the coarse-grained models we used the fine-grained model to guide us. For the WPI-5 model we started off knowing that we would have the 5 categories; 1) Algebra, 2) Geometry, 3) Data Analysis & Probability, 4) Number Sense and 5) Measurement. Both the National Council of Teachers of Mathematics and the Massachusetts Department of Education use these broad classifications as well as a 39 skill classification. After our 600 students had taken the 2005 state test, the state released the items from test and we had our subject matter expert tag up those test items.

The WPI-1, WPI-5 and WPI-39 models are derived from the WPI-106 model by nesting a group of fine-grained skills into a single category. This mapping is an aggregate or “is a part of” type of hierarchy as opposed to a prerequisite hierarchy [1]. Figure 1 shows the hierarchal nature of the relationship between WPI-106, WPI39, WPI-5 and WPI-1.

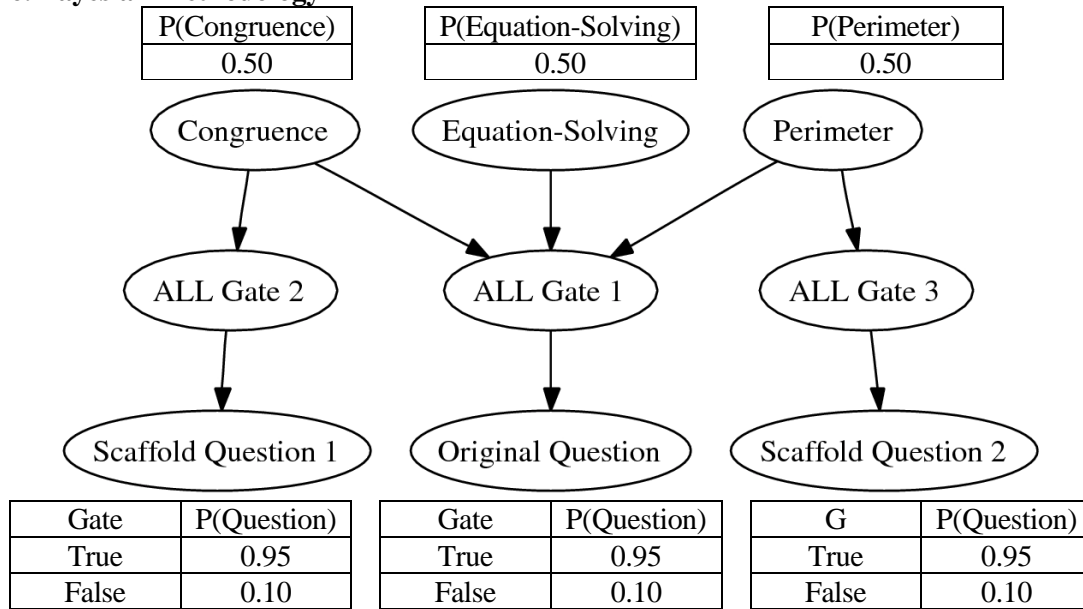
WPI-106	WPI-39	WPI-5	WPI-1
Inequality-solving Equation-Solving Equation-concept	setting-up-and-solving-equations	Patterns- Relations- Algebra	The skill of “math”
Plot Graph	modeling-covariation		
X-Y-Graph Slope	understanding-line-slope-concept		

Figure 1. Skill model hierarchy sample

#### 5. The Dataset

Both methods used the same data to evaluate users. Both methods used the same skill tagging as prescribed by the various grained skill models. The mixed-effects execution was memory limited and thus only a subset of 447, out of the 600 total users could be run. The Bayesian method used these same 447 users. This is the single point of difference between a previous Bayesian result [8] which ran 600 users. The current mixed-effects result also differs from a previous result [4] in the number of users run as well as the inclusion of the WPI-39 in this paper.

## 6. Bayesian Methodology



**Figure 2.** Bayesian Belief Network

The Bayesian topology for the 4 skill networks consist of skill nodes, ‘ALL’ nodes (which can be thought of as ‘AND’ nodes) and question nodes (See Figure 4). Each question node represents an original item or scaffold in the ASSISTment system and has its own ‘ALL’ node which has mapped to it the skills associated with that question according to the skill model. The reason for the ‘ALL’ node is to simplify the parameters for each question to only a guess and slip value which are set intuitively to 0.10 and 0.05 respectively. These values are what might be expected for a text-entry question. The parameters have not been optimized or allowed to vary per question or per model. The ‘ALL’ node also signifies that all the parent skills must be known in order for the questions to be answered correctly. The background probability of knowing a given skill is set to 0.50.

The prediction processes is run for one user at a time. The user’s data responses on the ASSISTment system are organized and entered into the Bayes net as evidence. The knowledge probabilities of the skills are then inferred. Now that we have predicted the user’s skills we can predict the test items. This is done by entering the inferred skill probabilities as “soft evidence” into the MCAS test network. Soft evidence is probabilistic as opposed to observed evidence. Now that the MCAS test network has skill values, we can infer the likelihood that a given question on the test will be answered correct. If the probability of correct for a question is inferred to be 0.70, then 0.70 points are added to the total score. Once all questions have been evaluated, a total projected MCAS score is left as a sum of the predicted question probabilities. This sum is compared with the user’s actual score for accuracy measures.

## 7. Mixed Effects method

For dichotomous (binary in our case) response data, several approaches adopting either a logistic or probit regression model and various methods for incorporating and estimating the influence of the random effects have been developed. Snijders & Bosker [13] provide a practical summary of the mixed-effects (fixed effect plus random effect) logistic regression model and various procedures for estimating its parameters. Hedeker & Gibbons [7] describes mixed-effects models for binary data that accommodate multiple random effects. As these sources indicate, the mixed-effects logistic regression model is a very popular and widely accepted choice for analysis of dichotomous data. It describes the relationship between a binary or dichotomous outcome and a set of explanatory variables. In this work, we adopted this model and fitted on our longitudinal, binary response data. When fitting the model, two sub-models will be simultaneously built, in which level-1 sub-model fits within-person change and describes how individuals change over time and level-2 sub-model tracks between-person change and describes how these changes vary across individuals. Such a model is often referred to as “longitudinal model” [12] since time is introduced as a predictor of the response variable, which allows us to investigate change over time.

After the model was constructed, the fixed-effects for the whole group and the random effects for each student were extracted and then the two learning parameters “intercept” and “slope” was calculated for each individual student (and for each skill if skill was introduced as factor into the model). Given this, we thus can apply the model on the items in the state test to estimate students’ response to each of them.

## 8. Comparing Test Results

For consistency, all models' logged results were arranged in standard format and evaluated by the same process to produce the average MAD and Error numbers in Table 1. The MAD score is the mean absolute difference which is the average difference between actual and predicted score across all users for a given model. MAD Error is the MAD score divided by the total number of questions on the test (MAD / 34). The under/over prediction is our predicted average score minus the actual average score on the test. The actual average score will be the same for all models. The centering is a result of offsetting every user’s predicted score by the average under/over prediction amount for that model and recalculating MAD and error percentage. This table of results appears in a poster paper at the AIED07 main conference [9].

**Table 1.** Bayesian and Mixed Effects Test Prediction Results

Model		MAD Error	MAD Score	Under/Over Prediction	Error After Centering	Centered MAD Score
WPI-106	Bayes	13.75%	4.19	1.0	13.60%	4.10
	Mixed-effects	12.10%	4.12	0.6	12.04%	4.10
WPI-39	Bayes	12.05%	4.10	1.0	11.82%	4.02
	Mixed-effects	12.40%	4.22	1.2	12.07%	4.11
WPI-5	Bayes	18.70%	5.42	3.5	16.20%	4.70
	Mixed-effects	12.84%	4.37	1.8	12.13%	4.12
WPI-1	Bayes	25.46%	7.39	4.6	21.63%	6.27

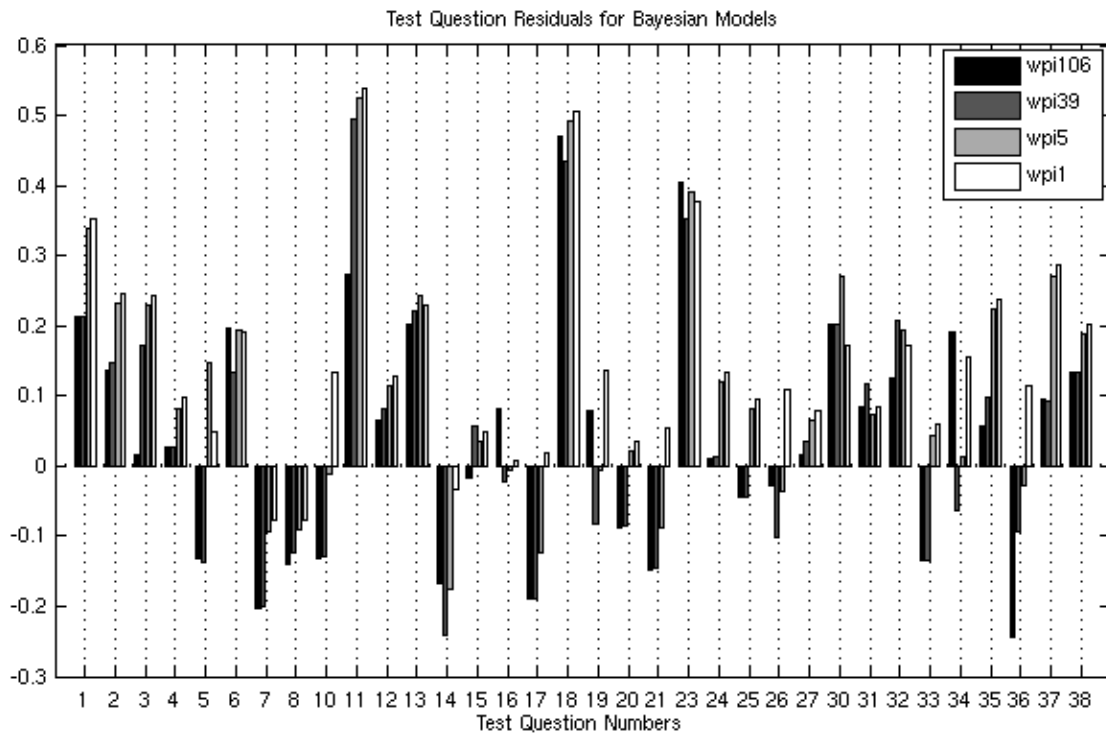
	Mixed-effects	13.00%	4.42	2.1	12.06%	4.10
<b>Best Bayes Error:</b> 12.05% (WPI-39)		<b>Best Mixed-effects Error:</b> 12.10% (WPI-106)				

We can see that the mixed-effects models outperform all but the Bayesian WPI-39 which stands as the highest performing model. A sizeable decrease in accuracy can be observed between the mixed-effects WPI-5 and WPI-1 and also in the Bayesian WPI-5 and WPI-1. This result suggests that the current Bayesian prediction performs best with the finer grained models. The top performing models for both methods are the fine-grained WPI-39 and WPI-106. The relative performance of the models are the same between the two methods with the exception of the WPI-39 and WPI-106. The paired T-test values of 0.8282 (Bayesian) and 0.2146 (mixed-effects) for the WPI-106 vs. the WPI-39 explains that the difference between the two models is not statistically significant and thus the two models are susceptible to variability amongst each other. All other models compared to one another are statistically significantly different in both methods ( $p < 0.05$ ).

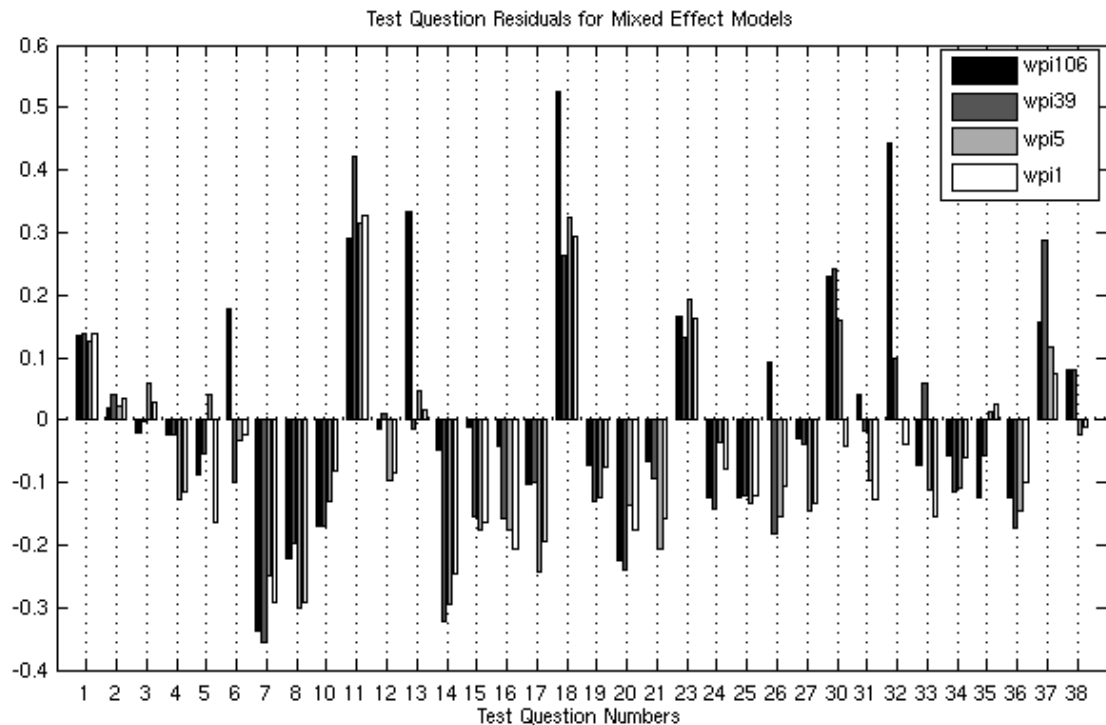
The internal fit of the models has also been evaluated using Bayesian methods [8]. The result was that number of skills in the model was proportionate to the accuracy of prediction. The WPI-106 did best at predicting the online student answers with 5.50% error while WPI-1 did the worst with 23.77% error. The methodology for predicting internal fit with the Bayes method is similar to the methodology described above for predicting the test. The difference in the training phase is that for each question predicted, the Bayes model was retrained using all the student data except the question being predicted. This was done for every question with every student. The error is the average percent difference between predicted total correct answers and actual correct answers. Internal fit was also evaluated with the Mixed-effects method. Though the evaluating methodology is different what we did with the Bayes method, the relative ranking of the modeling accuracy was the same: the more skills in the model, the better the internal fit. In addition, the result of the paired t-test showed that when mixed-effects method was used, finer grain-sized models always made statistically significantly better prediction on students' response in the online data.

## 9. Analysis of Residuals

To identify where the models struggle, we first look at the Bayesian test item residuals to find which questions have the lowest prediction accuracy. Figure 3 shows the Bayesian residuals for each model grouped by test question. Figure 4 shows the mixed-effects models' residuals. The test questions correspond to the question numbers of the 2005 MCAS 8<sup>th</sup> grade math test. Some numbers, such as 9, are skipped because they are short essay questions which we choose not to evaluate. A positive residual means that the item was underpredicted on average, a negative residual means the item was overpredicted. The residual is the average of the students' actual responses minus our predicted probabilities of correct for that question.



**Figure 3.** Bayesian residuals



**Figure 4.** Mixed-effects residuals

One observation that is immediately apparent in both the Bayes and mixed-effects models is that the residuals do not differ greatly per question from one skill model to the next. It is also apparent that the Bayes and mixed-effects residuals are quite similar. This similarity raises our confidence in the execution of the methodology used with the two models. A few differences we can point out between the two figures is that the WPI-106 model has the largest positive residual spikes in the mixed-effects model despite being the best performer. The

Bayesian wpi106 does not have these spikes. Also, question number 24 is an item of contention between the two methods with all the mixed-effects models overpredicting and all the Bayesian models underpredicting answers to the question. However, the two figures show decidedly similar traits best emphasized by questions 18 and 11 which are the worse predicted questions in both the Bayes and mixed-effects models. Question 18 is tagged with Venn-Diagram while question 11 is tagged with Point-Plotting. In the next section will investigate our online system to see if poor assessment of these skills is to blame.

## 10. Analysis of Online Skills: A case study of Venn-Diagram

A few items on the test stick out as being predicted very poorly. The reason for this could be a discrepancy between the knowledge displayed on the test for the skills relating to those questions and the assessed probability of knowledge for those skills after training on the online data. It could also be a poor tagging of skills to the questions. The former will be investigated in this section.

We will perform a case study of the skill of Venn-Diagram tagged to item number 18 on the test (Picture 1) which was the item with the highest average residual value among all the models. This item was consistently underpredicted on the test by about 40%. Our system believes this question should be a lot harder than it was, to answer why we gather information about that skill on our system. This information is shown in Table 2.

**Table 2.** Online skill information for Venn-Diagram

<b>Percent-Correct:</b>	18.2%
<b>Bayes Assesment Avg:</b>	22.9%
<b>Bayes Assesment Min:</b>	< .01%
<b>Bayes Assesment Max:</b>	99.8%
<b>Original Items:</b>	2
<b>Scaffold Items:</b>	4
<b>Data Points:</b>	2,106

**18** Coach Wilson constructed a Venn diagram that shows the number of eighth-grade athletes who play football, basketball, and hockey.

Which phrase best identifies the number 5 shown in the diagram?

- the total number of athletes who do not play all three sports
- the total number of athletes who play both football and hockey, but not basketball
- the total number of athletes who play either football or hockey, but not both
- the total number of athletes who do not play basketball

**Picture 1. MCAS Test Item 18**

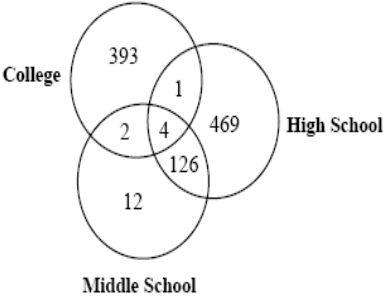
The percent correct of skills tagged with Venn-Diagram is 18.2%. This is a dramatic difference from the percent correct of 87.3% on the test question tagged with the same skill. This certainly suggests that the Venn-Diagram question in our online system is much harder than the question that appears on the test.

A significant difference between the two is that the online question is text entry where as the question on the test is multiple-choice. This difference in answer type creates a dramatic disparity in difficulty of the problem. We take a quick look at common answers for this online item in order to emphasize the relative ease of a multiple-choice questions on the MCAS test compared with their text entry counter parts in the ASSISTment system. The online problem was taken from the 2002 MCAS test (item 30) and converted to an ASSISTment item by removing the multiple-choices and adding scaffolding (Picture 2). There were ~1,200

responses to this question and the most common answer, with 455 responses, was the incorrect answer of 126 compared to the correct answer of 130 which received only 319 responses. The relevance to difficulty is that the common wrong answer of “126” was not an option among the 4 choices for the problem on the 2002 MCAS test which resulted in a 70% correct rate for that

30 The Venn diagram below shows Leila’s graduating classes from middle school, high school, and college.

**Leila’s Graduating Classes**



How many students graduated together from both Leila’s middle school and high school?

**Picture 2. ASSISTment Item**

question on the 2002 test with an IRT difficulty value of -1.4 (MCAS tech report). We can conclude that the multiple-choice question type was a significant factor in the poor predictive performance on that item. A more in depth analysis of buggy answers using this same item as an example can be found in a poster submission [11].

How can we correct for this disparity in difficulty between our online content and the

test? One approach is to optimize the parameters of the Bayes net. Right now the Bayesian online network does not make a distinction with regard to question type. This means that the same guess and slip parameters are used for multiple-choice and text entry questions. To address this, two separate guess and slip parameters were created to represent each question type. This was done using equivalence classes in BNT. Our current guess and slips values were set *ad hoc*, an improvement could be made if these values were learned from the data. To learn the guess and slip values for multiple-choice and text entry questions we use the Expectation Maximization algorithm. Both skill priors and guess/slip parameters were allowed to be optimized. An increase in accuracy of 0.77% or 0.22 MAD was attained using the optimized guess and slip parameters. This result was statistically significantly different than the result with unoptimized parameters. To further increase the performance of the network each question’s guess and slip parameter could be learned and hold out data could be used to train the parameters for the MCAS test network.

## 11. Conclusions

We have seen how the methods of Bayesian networks and mixed-effects modeling produce very similar results. This gives us confidence in the execution of the two methods and provides a rare doubly reinforced result arrived at from two different angles. We have also answered our research question about the utility of finer-grained models and can report that the fine-grained WPI-39 and WPI-106 models provide the most accurate test prediction as confirmed by both Bayesian and mixed-effects methods. Teachers will be happy to know that there is some validity to the fine-grained models. We also have shown how inspection of the models’ residuals can lead to a better understanding of transfer, content and the tutoring system as a whole.

There is one implication we would like to discuss. In the United States many schools and teachers are being encouraged to use frequent (i.e., monthly) testing to be “data-driven”. The problem is that many tests used are unidimensional and do not provide cognitive reporting

to the teachers while at the same time taking up valuable class time. There seems to be a tension between tests that are fast and tests that are cognitively diagnostic. One implication of the success of fine grained model assessment is that it might be possible for states to use these models to develop a system of their own similar to ASSISTment that does all three of these things; 1) accurately assesses students, 2) gives finer grained feedback that tends to be more cognitively diagnostic and 3) saves time by assessing students while they are getting “tutoring” or taking a test.

## 12. Future work

Our static analysis in this paper has shown encouraging results. In addition to grouped parameter learning we are also working on adding time to our Bayesian Network model. A preliminary evaluation with our temporal WPI-5 has shown a prediction improvement of 3.6% over the non temporal WPI-5. We also plan to evaluate other years’ datasets with both Bayesian and Mixed-effects methods. The results of those evaluations will tell if our results generalize beyond this particular year’s dataset.

## Acknowledgements

We would like to thank all the people associated with creating the Assistment system listed at [www.assistment.org](http://www.assistment.org) including the investigators Kenneth Koedinger, and Brian Junker at Carnegie Mellon. We would also like to acknowledge funding from the US Department of Education, the National Science Foundation, the Office of Naval Research and the Spencer Foundation. All of the opinions expressed in this paper are those solely of the authors and not those of our funders.

## References

- [1] Carmona, C., Millán, E., Pérez-de-la-Cruz, J.L., Trella, M. & Conejo, R. (2005) Introducing Prerequisite Relations in a Multi-layered Bayesian Student Model. In Ardissono, Brna & Mitrovic (Eds) User Modeling 2005; 10th International Conference. Springer. 347-356
- [2] J. Collins, J. Greer, and S. Huang. Adaptive assessment of using granularity hierarchies and Bayesian nets. In Proceedings of Intelligent Tutoring Systems, pages 569--577, 1996.
- [3] Fang Wei, Glenn D. Blank: Student Modeling with Atomic Bayesian Networks. Intelligent Tutoring Systems 2006: 491-502
- [4] Feng, M., Heffernan, N. T., Mani, M., & Heffernan, C. (2006). Using Mixed-Effects Modeling to Compare Different Grain-Sized Skill Models. In Beck, J., Aimeur, E., & Barnes, T. (Eds). Educational Data Mining: Papers from the AAAI Workshop. Menlo Park, CA: AAAI Press. pp. 57-66. Technical Report WS-06-05. ISBN 978-1-57735-287-7.
- [5] Feng, M., Heffernan, N.T., & Koedinger, K.R. (2006b). Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required. In Ikeda, Ashley & Chan (Eds.). Proceedings of the 8th International Conference on Intelligent Tutoring Systems. Springer-Verlag: Berlin. pp. 31-40. 2006.
- [6] Hambleton, R.K., & W. J. van der Linden. (1997). Handbook of modern item response theory. New York, NY: Springer-Verlag.
- [7] Hedeker, D. & Gibbons, Robert. D. (in progress). “Longitudinal Data Analysis”: “Mixed-Effects Regression Models for Binary Outcomes” (chapter 9).
- [8] Pardos, Z. A., Heffernan, N. T., & Anderson, B., Heffernan, C. L. The Effect of Model Granularity On Student Performance Prediction Using Bayesian Networks. 11th International Conference on User Modeling, 2007. Greece.
- [9] Pardos, Z. A., Feng, M. & Heffernan, N. T. & Heffernan-Lindquist, C. Analyzing fine-grained skill models using bayesian and mixed effect methods. In Luckin & Koedinger (Eds) Proceedings of the 13th Conference on Artificial Intelligence In Education. IOS Press.
- [10] Reye, J. (2004). Student modelling based on belief networks. International Journal of Artificial Intelligence in Education: Vol. 14, 63-96.
- [11] Rob Weitz, Neil Heffernan, Viswanathan Kodaganallur, David Rosenthal (submitted). The Distribution of Student Errors Across Schools: An Initial Study. AIED 2007
- [12] Singer, J. D. & Willett, J. B. (2003). Applied Longitudinal Data Analysis: Modeling Change and Occurrence. Oxford University Press, New York.
- [13] Snijders, Tom A. B., and Bosker, Roel J. (1999). Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, London etc.: Sage Publishers, 1999.