The Effect of Model Granularity on Student Performance Prediction Using Bayesian Networks[1]

Zachary A. Pardos, Neil T. Heffernan, Brigham Anderson, Cristina L. Heffernan

Worcester Polytechnic Institute, Carnegie Mellon University

{zpardos, nth, ch}@wpi.edu, brigham@cmu.edu

Abstract. A standing question in the field of Intelligent Tutoring Systems and User Modeling in general is what is the appropriate level of model granularity (how many skills to model) and how is that granularity derived? In this paper we will explore models with varying levels of skill generality (1, 5, 39 and 106 skill models) and measure the accuracy of these models by predicting student performance within our tutoring system called ASSISTment as well as their performance on a state standardized test. We employ the use of Bayesian networks to model user knowledge and for prediction of student responses. Our results show that the finer the granularity of the skill model, the better we can predict student performance within the tutor. However, for the standardized test data we received, it was the 39 skill model that performed the best. We view this as support for fine-grained skill models despite the finest grain model not predicting the state test scores the best.

[1] This paper is an expansion of work presented at the workshop in Educational Data Mining held at the 18th International Conference on Intelligent Tutoring Systems (2006) in Taiwan [13] and the 11th International Conference on User Modeling (2007) in Corfu, Greece [14]

1. Introduction

Most large standardized tests (such as the SAT or GRE) are what psychometricians call "unidimensional" in that they are analyzed as if all the questions are tapping a single underlying knowledge component (i.e., skill). However, cognitive scientists such as Anderson & Lebiere [1] believe that students are learning individual skills and might learn one skill but not another. Among the reasons psychometricians analyze large scale tests in a unidimensional manner is that student performance on different skills is usually highly correlated, even if there is no necessary perquisite relationship between these skills. We are engaged in an effort to investigate if we can do a better job of predicting student performance by modeling individual skills. We consider four different *skill models*[2]; one that is unidimensional, WPI-1, one that has five skills we call the WPI-5, one that has 39 skills called the WPI-39 and our most fine-grained model that has 106 skills which we call the WPI-106. We will refer to a tagging of skills to questions as a skill model. We will compare skill models that differ in the number of skills and see how well the different models can fit a data set of student responses collected via the ASSISTment system.

There are many researchers in the user modeling and educational data mining communities working with Intelligent Tutoring Systems who have adopted Bayesian network methods for modeling knowledge [3, 7, 12, 19]. Even methods that were not originally thought of as Bayesian Network methods turned out to be so; Reye [17] showed that the classic Corbett & Anderson "Knowledge tracing" approach [8] was a special case of a dynamic belief network.

[2] A *skill-model* is referred to as a "Q-matrix" by some AI researchers [5] and psychometricians [18]

We are not the first to do model-selection based on how well the model fits real student data [10, 12]. Nor are we the only ones that have been concerned with the question of granularity; Greer and colleagues [11] have proposed methods of assessment using different levels of granularity to conceptualize student knowledge. However, we are not aware of any other work where researchers attempted to specifically answer the question of "what is the right level of granularity to best fit a data set of student responses".

1.1 Background on the MCAS Test

The Massachusetts Comprehensive Assessment System (MCAS) is a Massachusetts state administered standardized test that covers English, math, science and social studies for grades 3rd through 10th. We focused on the 29 multiple choice questions of the 8th grade mathematics test only. Our work related to the MCAS in two ways. First, we built our tutor content based upon 300 publicly released items from previous MCAS math tests. Secondly, we evaluated our models by predicting students' scores on the 8th grade 2005 MCAS test which was taken at the end of the school year. The MCAS test is of particular importance to students and teachers in Massachusetts because students must pass the test in order to graduate high school.

1.2 Background on the ASSISTment tutor system

The ASSISTment system is an e-learning and e-assessing system. 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" to. Students get this tutoring, which we refer to as scaffolding, when they answer an original question incorrectly. 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 into parts (as shown in Figure 1) which allows us to tell if the student answered incorrectly because he or she did not know one skill versus another. Students answered 100 original questions and 160 scaffold questions on average. A student is only marked as getting the item correct if he or she answered the question correctly on the first attempt without assistance from the system. The 2005 MCAS test items were publically released in June of 2005. We tagged

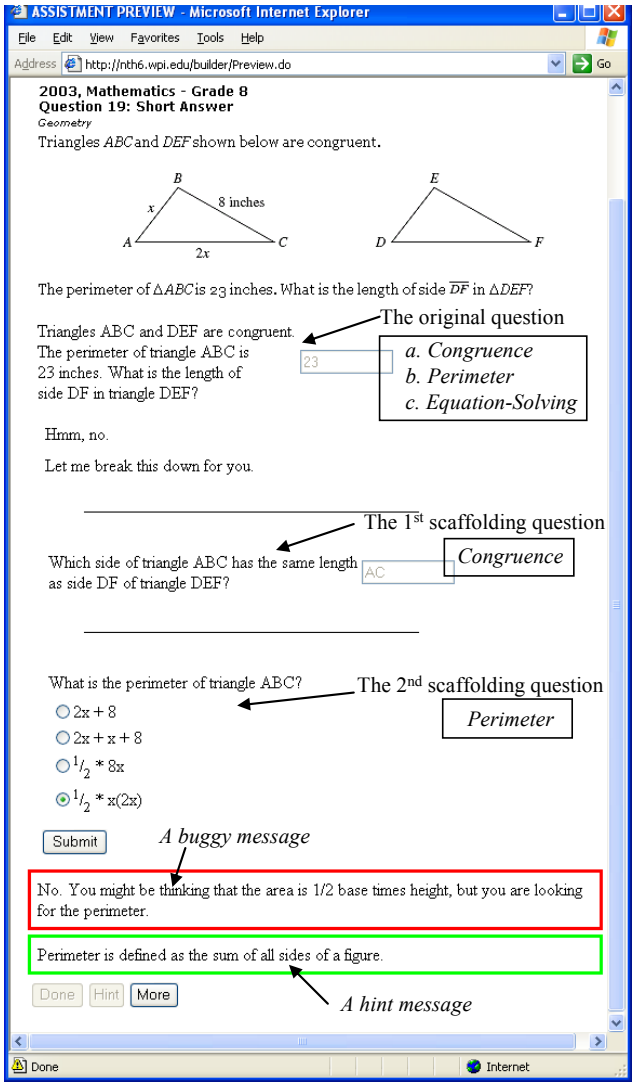
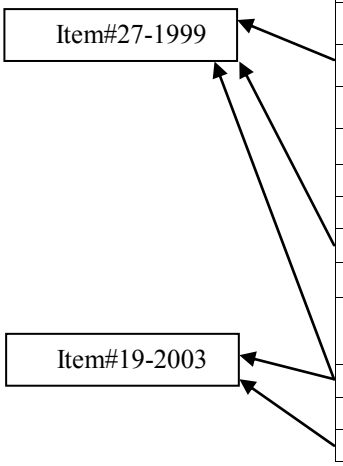


Figure 1. An ASSISTment showing the original question and the first two scaffolding

these items with skills shortly after they were released but before we received the students' official scores from the state.

2. Models: Creation of the Fine-Grained Skill Model

In April of 2005, a seven hour "coding session" was staged where our subject-matter expert, Cristina Heffernan, with the assistance of the 2nd author, set out to make up skills and tag all of the 300 existing 8th grade MCAS items with these skills. 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 imposed upon our subject-matter expert that no one item would be tagged with more than three skills. She gave the skills names, but the real essence of a skill is what items it was tagged to. This model is referred to as the 'April' model or the WPI-106. The National Council of Teachers of Mathematics and the Massachusetts Department of Education use broad classifications of five and 39 skill sets. The 39 and five skill classifications were not tagged to the questions. Instead, the skills in the coarse-grained models were mapped to the finest-grained model in a "is a part of" type of hierarchy, as opposed to a prerequisite hierarchy [6]. The appropriate question-skill tagging for the WPI-5 and WPI-39 models could therefore be derived from this hierarchy as illustrated in Figure 2, bellow.



WPI-106	WPI-39	WPI-5	WPI-1
Inequality-solving	setting-up-and-solving-equations	Patterns-Relations-Algebra	The skill of "math"
Equation-Solving			
Equation-concept			
Plot Graph	modeling-covariation		
X-Y-Graph	understanding-line-slope-concept		
Slope			
Congruence	understanding-and-applying-congruence-and-similarity	Geometry	
Similar Triangles			
. . .	. . .	. . .	
Perimeter	using-measurement-formulas-and-techniques	Measurement	
Circumference			
Area			

Figure 2. Questions are tagged with the WPI-106 which is mapped to the other skill models

2.1 How the Skill Mapping Was Used to Create a Bayesian Network

Our Bayesian Networks consisted of three layers of binomial random variable nodes as illustrated in Figure 3. A separate network was created for each skill model. The top layer nodes represent knowledge of a skill that was set to a prior probability of 0.50. This model is simple and assumes all skills are as equally likely to be known prior to being given any evidence of student responses, but once we present the network with evidence it can quickly infer probabilities about what the student knows. The bottom layer nodes are the question nodes with conditional probabilities set *ad-hoc* to 0.10 for guess and 0.05 for slip. The intermediary 2nd layer consists of AND[3] gates that, in part, allowed us to only specify a guess and slip parameter for the question nodes regardless of how many skills were tagged to them. Our colleagues [2] investigated using a compensatory model with the same dataset but we found [15] that a conjunctive, AND, is very well suited to model the composition of multiple

[3] An 'ALL' gate is equivalent to a logical AND. Kevin Murphy's Bayes Net Toolbox (BNT) evaluates MATLAB's ALL function to represent the Boolean node. This function takes a vector of values as opposed to only two values if using the AND function. Since a question node may have three skills tagged to it, the ALL function is used.

skills. When predicting MCAS test questions, a guess value of 0.25 was used to reflect the fact that the MCAS items being predicted were all multiple choice (out of four), while most of the online ASSISTment items have text-input fields as the answer type. Future research will explore learning the parameters from data.

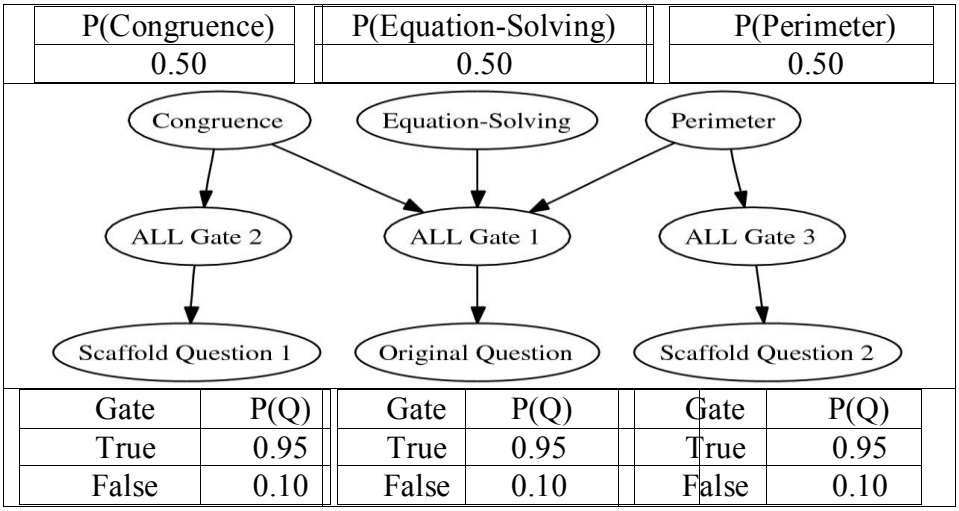


Figure 3. Example of the Bayesian network topology for an ASSISTment with two scaffolds. P(Q=Correct|Gate=False) represents the slip. 1 – P(Q=Correct|Gate=True) represents the slip.

2.2 Model Prediction Procedure

A prediction evaluation was run for each model one student at a time. The student's responses on the tutor were presented to the Bayesian network as evidence and inference (using exact join-tree) was made on the skills to attain knowledge probabilities. To predict each of the 29 questions we used the inferred skill probabilities to ask the Bayesian network what the probability was that the student would get the test question correct. We calculated a *predicted score* by taking the sum of the probabilities for all test questions. Finally, we found the

percent error by taking the absolute value of the difference between predicted and actual score and dividing that by 29. The *Average Error* of a skill model is the average error across the 600 students. Table 1, bellow, demonstrates how the error was calculated. Notice that the predicted probability of answering correctly, P(q), is the same for test questions of the same skill (questions 1 and 2 for example). Also notice test question 3 involves two skills, Patterns and Measurement, and the P(q) of that question is lower than the P(q) of questions with only one of either skill because of the conjunctive (AND) model of skill composition being used.

Test Question	**Skill Tagging (WPI-5)**	**user 1 P(q)**	**user 2 P(q)**	**…**	**user 600 P(q)**	**Average Error**
1	Patterns	0.2	0.9	…	0.4	
2	Patterns	0.2	0.9	…	0.4	
3	Patterns & Measurement	0.1	0.5	…	0.2	
4	Measurement	0.8	0.8	…	0.3	
5	Patterns	0.2	0.9	…	0.4	
:	:	:	:	:	:	
29	Geometry	0.7	0.7	…	0.2	
	Predicted Score	14.2	27.8	…	5.45	
	Actual Score	18	23	…	9	
	Error	10.34%	19.42%	…	12.24%	**17.28%**

Table 1. Tabular illustration of error calculation

3. Results

An early version of the results in this section (using approximate inference instead of exact inference and without Section 3.1) appears in a workshop paper [13]. The MAD score is the mean absolute difference between predicted and actual score. The under/over prediction is our predicted average score minus the actual average score on the test.

Table 2. Model prediction performance results for the MCAS test

Model	Error	MAD	Under/Over
WPI-39	12.86%	3.73	↓ 1.4
WPI-106	14.45%	4.19	↓ 1.2
WPI-5	17.28%	5.01	↓ 3.6
WPI-1	22.31%	6.47	↓ 4.3

The results in Table 2, above, show that the WPI-39 had the best accuracy with an error of 12.86% which translates to a raw score error of 3.73. The finest-grain model, the WPI-106, came in second followed by the WPI-5 and finally the WPI-1. We can conclude that the fine grain models are best for predicting the external test but that the finest grain model is not number one. An investigation [16] of error residuals revealed that test questions that were poorly predicted had a dramatically higher percent correct on the test than questions on the ASSISTment system relating to the same skill. These ASSISTment questions all had text input fields (fill in the blank) question types. The conclusion drawn was that the multiple choice question type of the test made some questions much easier than their ASSISTment counter parts to an extent that was not captured by the guess and slip of the model. Learning the guess and slip parameters might correct for the consistent under predicting shown in the results. All results in Table 2 were statistically significantly separable from each other at the $p < .05$ level.

3.1 Internal/Online Data Prediction Results

To answer the research question of which level of granularity is best for predicting student performance *within the system* we measure the internal fit. The internal fit is how accurately

we can predict student answers to our online question items. If we are able to accurately predict a student's response to a given question, this brings us closer to a computer adaptive tutoring application of being able to intelligently select the appropriate next questions for learning and or assessing purposes.

The internal fit prediction was run, again, for one student at a time. The processes was similar to an N-fold cross validation where N is the number of question responses for that student. The network was presented with evidence minus the question being predicted. One point was added to the internal total score if the probability of correct was greater than 0.50 for the question. This was repeated for each question answered by the student. The mean absolute difference between predicted total and actual total score was tabulated in the same fashion as the results of Table 1 however it is worthwhile to note that not all users answered the same number of questions so the MAD will have greater variance. All the differences between the models in Table 3 were statistically significantly different at the $p < .05$ level.

Table 3. Model prediction performance results for internal fit

Model	Error	MAD	Under/Over
WPI-106	5.50%	15.25	↓ 12.31
WPI-39	9.56%	26.70	↓ 20.14
WPI-5	17.04%	45.15	↓ 31.60
WPI-1	26.86%	69.92	↓ 42.17

We can see from the results above that prediction accuracy increase at a greater than linear rate with the number of skills in the skill model. For predicting student performance on the tutor system, the finer grained the model, the better. We can also observe under prediction of

all models which again suggests there is room for improvement by learning better parameter values for the network.

4. Discussion and Conclusions

It appears that we have found good evidence that fine-grained models can produce better tracking of student performance as measured by ability to predict student performance on a state test and the tutor. The WPI-39 was found to be the most accurate at state test prediction. We are glad to see that by paying attention to granularity we can do a better job, however, the finest grained model was expected to be the best predictor. One possible explanation for why it was not is that the 29 question test represents a small subset of the 109 skills. So the WPI-106 is left at a disadvantage since only 27% of the skills it is tracking appear on the 2005 MCAS test. Essentially 75% of the data that the WPI-106 collects is thrown out resulting in less data per skill compared to other models. The WPI-39, on the other hand, can benefit from its relatively fine-grained tracking and 46% of its skills are sampled on the 29 item MCAS test. A general consideration is that the finer grained the model, the more data is required to sufficiently sample all skills and also the more content that is required to represent each skill. The internal fit however showed that the finer grained the model, the better the fit to the data collected from the ASSISTment system. So, with systems that collect lots of data, finer grained skill models excel in student assessment and performance prediction. This result is in accord with other work we have done using mixed-effect-modeling [9, 16].

Some of our colleagues have perused item response models for this very dataset [2, 4] with considerable success, but we think that item response models don't help teachers

identify what skills a students should work on, so even though it might be very good predictor of students, it suffers in other ways. Part of the utility of fine-grained modeling is being able to identify skills that students have mastered and those that need work. An example of a class skill report presented to a teacher is shown in Figure 4, bellow.

Top 5 hard knowledge components

WPI 8th Grade Math Fine Grained Model Click to sort by	Skill Meter	Rate Click to sort by	#Record Click to sort by
Surface-Area		8%	35
Congruence		18%	16
inverse relations		31%	32
Pythagorean-theorem		33%	54
Division-Fraction		42%	21

Figure 4. Class skill report generated by the WPI-106 fine-grained skill model

We think that this work is important, in that while adapting fine-grained models is hard, we have shown they can result in better prediction of important measures such as state test scores. There are still several good reasons for psychometricians to stick with their uni-dimensional models, such as the fact that most tests have a small number of items, and they don't have scaffolding questions that can help deal with the hard credit-blame assignment problems implicit in allowing multi-mapping (allowing a single question to be tagged with more than one skill).

Acknowledgements. This research was made possible by the US Dept of Education, Institute of Education Science, "Effective Mathematics Education Research" program grant #R305K03140, the Office of Naval Research grant #N00014-03-1-0221, NSF CAREER award to Neil Heffernan, and the Spencer Foundation. All of the opinions in this article are

those of the authors, and not those of any of the funders. This work would not have been possible without the assistance of the 2004-2005 WPI/CMU ASSISTment Team that helped make possible this dataset. The first author is an NSF GK-12 fellow.

REFERENCES

[1] Anderson, J. R. & Lebiere, C. (1998). The Atomic Components of Thought. LEA.

[2] Anozie N., & Junker B. W. (2006). Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system**.** In Beck, J., Aimeur, E., & Barnes, T. (Eds). Educational Data Mining: Papers from the AAAI Workshop. Menlo Park, CA: AAAI Press. pp. 1-6. Technical Report WS-06-05.

[3] Arroyo, I., Woolf, B. (2005) Inferring learning and attitudes from a Bayesian Network of log file data. Proceedings of the 12thInternational Conference on Artificial Intelligence in Education. 33-40.

[4] Ayers E., & Junker B. W. (2006). Do skills combine additively to predict task difficulty in eighth grade mathematics? In Beck, J., Aimeur, E., & Barnes, T. (Eds). Educational Data Mining: Papers from the AAAI Workshop. Menlo Park, CA: AAAI Press. pp. 14-20. Technical Report WS-06-05.

[5] Barnes, T. (2005), Q-matrix Method: Mining Student Response Data for Knowledge. In the Technical Report (WS-05-02) of the AAAI-05 Workshop on Educational Data Mining, Pittsburgh, 2005.

[6] Carmona, C., Millán, E., Pérez-de-la-Cruz, J.L., Trella1, M. & Conejo, R. (2005) Introducing Prerequisite Relations in a Multi-layered Bayesian Student Model.

In Ardissono, Brna & Mitroivc (Eds) *User Modeling 2005; 10th Internaton Confrence.* Springer. 347-356

[7] Conati, C., Gertner, A., & VanLehn, K. (2002). Using bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction*, *12*(4), 371–417.

[8] Corbett, A. T., Anderson, J. R. & O'Brien, A. T. (1995) Student modeling in the ACT programming tutor. Chapter 2 in Nichols, P. D., Chipman, S. F. and Brennan, R. L. (eds.) (1995). Cognitively diagnostic assessment. Lawrence Erlbaum Associates: Hillsdale, NJ.

[9] 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.

[10] Mathan, S. & Koedinger, K. R. (2003). Recasting the Feedback Debate: Benefits of Tutoring Error Detection and Correction Skills. In Hoppe, Verdejo & Kay (Eds.), *Artificial Intelligence in Education: Shaping the Future of Learning through Intelligent Technologies., Proceedings of AI-ED 2003* (pp. 39-46). Amsterdam, IOS Press.

[11] McCalla, G. I. and Greer, J. E. (1994). Granularity-- based reasoning and belief revision in student models. In Greer, J. E. and McCalla, G. I., editors, Student Modelling: The Key to Individualized Knowledge--Based Instruction, pages 39--62. Springer--Verlag, Berlin.

[12] Mislevy, R.J., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. User-Modeling and User Adapted Interaction, 5, 253-282.

[13] Pardos, Z. A., Heffernan, N. T., & Anderson, B., Heffernan, C. L. (2006) Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks. Workshop in

Educational Data Mining held at the Eight International Conference on Intelligent Tutoring Systems. Taiwan.

[14] Pardos, Z. A., Heffernan, N. T., Anderson, B. & Heffernan, C. (2007). *The effect of model granularity on student performance prediction using Bayesian networks*. Proceedings of the 11th International Conference on User Modeling. Corfu, Greece. Springer Berlin. pp. 435-439.

[15] Pardos, Z. A., Heffernan, N. T., Ruiz, C. & Beck, J. (2008). *The Composition Effect: Conjunctive or Compensatory? An Analysis of Multi-Skill Math Questions in ITS*. Proceedings of the First Conference on Educational Data Mining. Montreal, Canada. pp. 147-156.

[16] Pardos, Z. A., Feng, M. & Heffernan, N. T. & Heffernan-Lindquist, C. (2007) *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. pp. 626-628.

[17] Reye, J. (2004). Student modelling based on belief networks. *International Journal of Artificial Intelligence in Education: Vol. 14,* 63-96.

[18] Tatsuoka, K.K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto, (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 453-488). Hillsdale, NJ: Lawrence Erlbaum Associates.

[19] Zapata-Rivera, J-D and Greer, J.E. (2004). Interacting with Inspectable Bayesian Models. *International Journal of Artificial Intelligence in Education*. *Vol. 14,* 127-163.