Using Mixed-Effects Modeling to Analyze Different Grain-Sized Skill Models in an Intelligent Tutoring System

Mingyu Feng, Neil T. Heffernan, Cristina Heffernan, and Murali Mani

Abstract—Student modeling and cognitive diagnostic assessment are important issues that need to be addressed for the development and successful application of intelligent tutoring systems (ITS). ITS needs the construction of complex models to represent the skills that students are using and their knowledge states, and practitioners want cognitively diagnostic information at a finer grained level. Traditionally, most assessments treat all questions on the test as sampling a single underlying knowledge component. Can we have our cake and eat it, too? That is, can we have a good overall prediction of a high stakes test, while at the same time be able to tell teachers meaningful information about fine-grained knowledge components? In this paper, we introduce an online intelligent tutoring system that has been widely used. We then present some encouraging results about a fine-grained skill model with the system that is able to predict state test scores. This model allows the system track about 106 knowledge components for eighth grade math. In total, 921 eighth grade students were involved in the study. We show that our fine-grained model could improve prediction compared to other coarser grained models and an IRT-based model. We conclude that this intelligent tutoring system can be a good predictor of performance.

Index Terms—Intelligent tutoring systems, cognitive diagnostic assessment, fine-grained skill model, statistical analysis of skill models, mixed-effects model.

1 Introduction

An intelligent tutoring system usually consist of four different modules: the interface module where students interact with the ITS, the expert module that refers to an expert model containing description of knowledge the ITS is teaching, the student module that contains descriptions of student knowledge or behaviors, and the tutor module that takes corrective action such as providing feedback or remedial instruction. This work is concerned with the construction of better expert models. More precisely, our research question is what the right grain size of expert models (referred to as skill models in the paper) are and how model granularity impacts the effectiveness of student's knowledge tracking. This paper is based on the work we conducted within an intelligent tutoring system called the ASSISTment System, introduced below.

1.1 Instructional Assistance + Formative Assessment = ASSISTments

In many states, there are concerns about poor student performance on new high-stakes standards-based tests that are required by the No Child Left Behind Act (NCLB). Partly because of this pressure, and partly because teachers, parents, and other stakeholders want and need more immediate feedback about how students are doing, there is a growing interest in "Formative Assessment" in K-12 Education [33] with many companies[1] providing such services. Some teachers make extensive use of practice tests and released test items to help identify learning deficits for students. However, such tests not only require great effort and dedication, but they also take valuable time away from instruction. The limited classroom time available in middle school mathematics classes compels teachers to choose between time spent assisting students' development and time spent assessing students' abilities. A solution must involve a way whereby students can take an assessment and learn simultaneously. Yet, traditionally, these two areas of testing (i.e., Psychometrics) and instruction (i.e., math educational research and instructional technology research) have been separated fields of research with their own goals. Statisticians have not done a great deal of work to enable assessment of students while they are learning.[2] To help resolve this dilemma, the US Department of Education funded Heffernan and Koedinger to build a Web-based tutoring system ("the ASSISTment System"[3]) that would also do assessment at the same time.

- *M. Feng is with the Department of Computer Science, Worcester Polytechnic Institute, 100 Institute Road, WPI#696, Worcester, MA 01609. E-mail: mfeng@cs.wpi.edu.*
- *N.T. Heffernan is with the Department of Computer Science, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609. E-mail: nth@wpi.edu.*
- *C. Heffernan and M. Mani are with the Department of Computer Science, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609. E-mail: {ch, mmani}@cs.wpi.edu.*

Manuscript received 26 Dec. 2008; revised 24 Mar. 2009; accepted 3 Apr. 2009; published online 8 Apr. 2009.
For information on obtaining reprints of this article, please send e-mail to: lt@computer.org, and reference IEEECS Log Number TLTSI-2008-12-0122.
Digital Object Identifier no. 10.1109/TLT.2009.17.

1. Including assessment systems from Northwest Evaluation Association (http://nwea.org/assessments/), Measured Progress (http://measuredprogress.org), Pearson (http://www.pearsonassessments. com/), and the Center for Data-Driven Reform in Education (http://www.cddre.org/Services/4Sight.cfm).

2. Standard psychometric models assume that the amount of learning happens during a test is limited. Some works have been done to measure growth and change (e.g., [42], [16]), but they are not based on testing data where students are actively learning materials.

3. The term "ASSISTment" was coined by Kenneth Koedinger and blends instructional **assist**ance and assess**ment**.

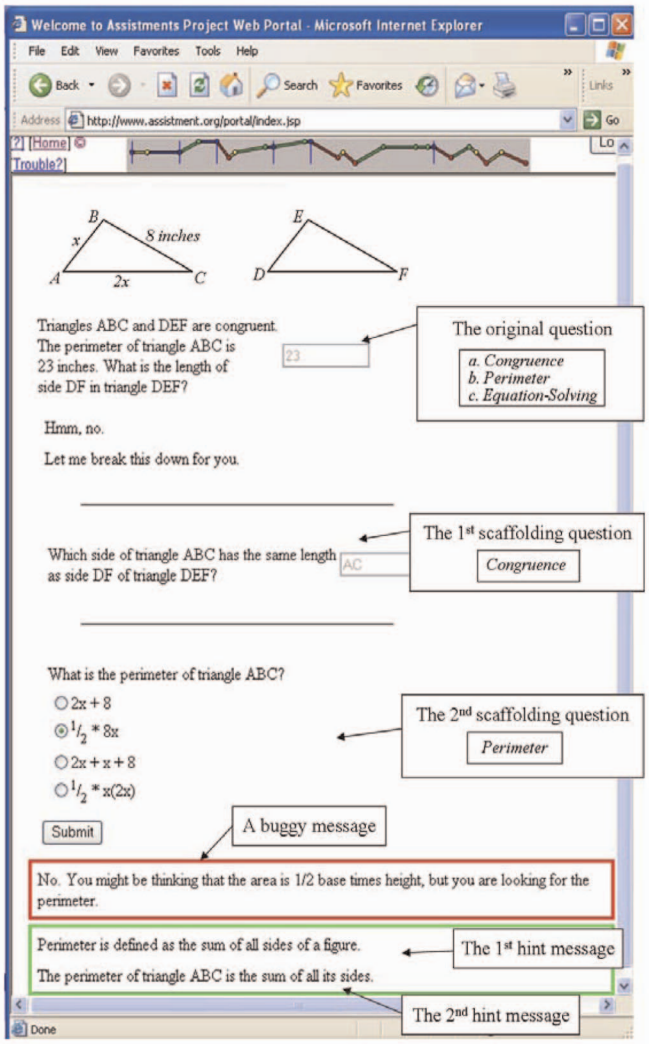


Fig. 1. An ASSISTment question showing a student working in the system.

In Massachusetts, the state department of education has released 11 years (1998-2008) worth of eighth grade Massachusetts Comprehensive Assessment System (MCAS) test items that have been turned into ASSISTment System by adding tutoring. Each ASSISTment question consists of an *original question* and a list of *scaffolding questions*. The original question usually has the same text as found in the MCAS test, while the scaffolding questions were created through breaking the original question down to the individual steps by our content experts. A student is initially presented a question that usually has several skills needed to solve it correctly. If the student gets the question correct, he would get credit for all the associated skills and move on to next question, otherwise he is forced to go through a sequence of scaffolding questions (or scaffolds). Students work through the scaffolding questions, possibly with hints and buggy messages, until they eventually get the problem solved. An ASSISTment question that was built for Item 19 of the 2003 MCAS is shown in Fig. 1. We see that the student typed "23," a wrong answer, for the original question that involved understanding algebra, perimeter, and congruence. Once the student gets the first scaffolding question correct (by typing "*AC*"), the second scaffolding question appears, focusing on the concept of perimeter. After he got this question right, he was given a question on equation-solving. Buggy messages will show up if the student types in a wrong answer. So, if a student got the original question wrong, what skills have they not mastered? A fine-grained skill model will help determine which of the skills needed to solve this problem that the student has not mastered.

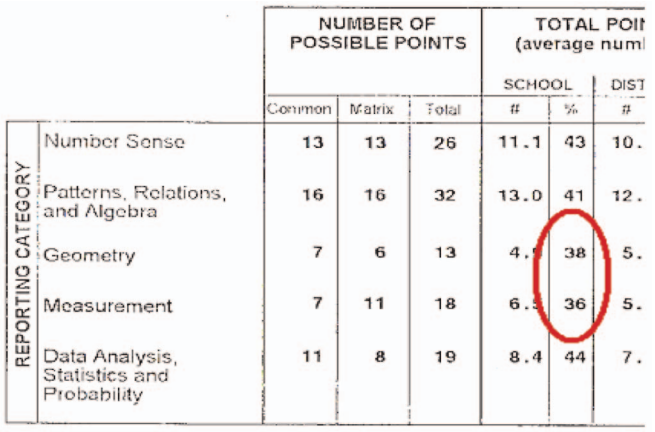
Since launched in 2004, the number of users of the ASSISTment System has been expanded every year. More than 3,000 students from Massachusetts and Pennsylvania used the system during the school year of 2007-2008 in their normal math class. A number of studies have been conducted to evaluate the impact of the ASSISTment System as an assessor and as a tutor. Our work [17], [22] showed more accurate assessment that can be achieved by not only using the overall correctness of student answer, but also using the interaction data, such as response speed, help-seeking behavior on the effort required for students to solve a test item with instructional assistance. It has been shown that students are learning from working in the ASSISTment System [21], [39], [40]. Additionally, randomized controlled experiments have been conducted to determine the effectiveness of different types of interventions [39], [40].

1.2 Motivation

Issues for practitioners. Most large standardized tests are "unidimensional" in that they are analyzed as if all the questions are tapping a single underlying skill. However, cognitive scientists such as Anderson and Lebiere [3] believe that students are learning individual skills. Among the reasons that psychometricians analyze large-scale tests in a unidimensional manner is that students' performance on different skills are usually highly correlated, even if there is no necessary prerequisite relationship between these skills. Another reason is that students usually do a small number of items in a given setting (for instance, 39 items for the eighth grade math MCAS test), which makes it hard to acquire identifiability for each single skill, especially when the number of skills that need to be mastered is larger than the number of the items in the test. Such tests work pretty well at telling you which students are performing well but are not good at *informing educators* about which skills are causing difficulty and how to help students.

However, the question of tagging items to learning standards is very important because schools seek to use the MCAS assessments in a data-driven manner to provide regular and ongoing feedback to teachers and students on progress toward instructional objectives. However, while the number of mathematics skills and concepts that a student needs to acquire is on the order of hundreds, the feedback on the MCAS to principals, teachers, parents, and students is broken down into only five mathematical reporting categories, known as "Strands." And the state's "Curriculum Framework" breaks the five strands into 39 individual "learning standards" for eighth grade math and tags each item with one of the 39 standards. The MCAS reporting system is representative of other states' reporting systems.

In 2004, a principal handed us a report (shown in Fig. 2) he received from the state and asked that we focus efforts



		NUMBER OF POSSIBLE POINTS			TOTAL POI (average numl		
					SCHOOL		DIST
		Common	Matrix	Total	#	%	#
REPORTING CATEGORY	Number Sense	13	13	26	11.1	43	10.
	Patterns, Relations, and Algebra	16	16	32	13.0	41	12.
	Geometry	7	6	13	4.	38	5.
	Measurement	7	11	18	6.	36	5.
	Data Analysis, Statistics and Probability	11	8	19	8.4	44	7.

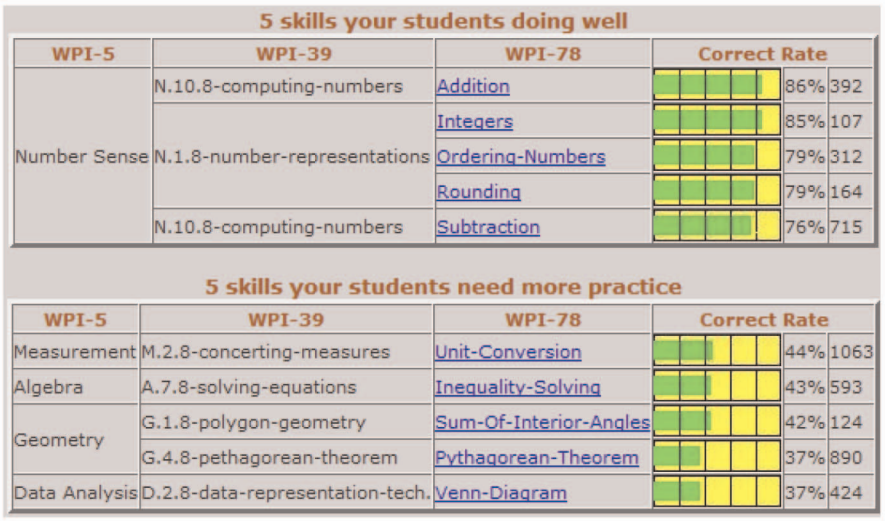
Fig. 2. A school-level report showing low percent correct for geometry and measurement on the MCAS Test.

on Geometry and Measurement because his students scored poorly in those areas (receiving 38 and 36 percent correct compared to over 41+ percent correct in the three other reporting categories). However, a detailed analysis of state tests in Texas concluded that such topic reporting is not reliable because items are not equated for difficulty within these areas (Confrey et al. 2002). Receivers of such reports are being told to be "data-driven" and use the reports to inform their instruction yet the reports themselves are never designed to give feedback at a grain size that could be used for this purpose. A reader can get some intuition on why this is the case by trying item 19 from the 2003 MCAS shown in Fig. 1. Then ask yourself, "What makes this item difficult?" Clearly, this item includes elements from four of the five strands: Algebra, Geometry (for its use of congruence), Number Sense (for doing the arithmetic operations), or Measurement (for the use of perimeter). Yet, the state chose just one of the five strands, Geometry, to classify the item ignoring this obvious overlap. When we asked this same question about item 19 to teachers on our Teacher Advisory Board, one of our cooperating educators said "But you know the state—I worked on the MCAS assessment for perimeter [referring to one of the committees that designs items for the MCAS] . . . and sometimes when they get a problem that they know very well that it could fit more than one strand, they put it in the strand that *they* need to fill. And that is the problem!" Another teacher followed up with "It does affect reports . . . because then the state sends reports that say that your kids got this problem wrong so they are bad in geometry—and you have no idea, well you do not know what it really is—whether it is algebra, measurement/perimeter, or geometry." Thus, a teacher cannot trust that putting more effort on a particular low scoring area will indeed pay off in the next round of testing. It was reported that instead of having performance reports that break math knowledge into only a few components, teachers want more fine-grained diagnostic reports to accommodate their everyday classroom practice. These reports are referred to as "assessment for learning" (e.g., [29], [45]).

Needs of intelligent tutoring systems. One key component of creating an intelligent tutoring system is forming the model that monitors student behavior. ITS needs the construction of complex models to represent the skills that students are using and their knowledge states. As students work through the program, the model tracks their progress and chooses what problems will be displayed next. By using a better skill model, a system should be able to do a better job of predicting which items students will get correct in real time. That means that the system can do a better job of selecting the next best item for students to work on. For instance, one criterion of the next "best" item could be the one that has the largest ratio of expected test score gain to expected time to complete the problem. Expected test score gain will be a function that depends upon both the expected rise in skills from doing that item at that time, as well as the weight of those skills on the test (i.e., the MCAS). A better model would also help in addressing the issues that we mentioned above to help teachers adjust their instruction in a data-driven manner. Such a model will allow a teacher who has one week before the MCAS to know what topics to review to maximize the class average. We can make a calculation averaging the whole class to suggest what will give the teacher the biggest "bang for the buck." An example of a useful report [19] that teachers can get using the ASSISTment system is shown in Fig. 3. Teachers can see how their students are doing on each skill and can determine where they need to spend the most time.

We are engaged in an effort to investigate if we can do a better job of predicting the state test scores by modeling individual skills in a finer grain size. This is not applicable in the traditional assessment environment because of limited testing time and test items. It is hard to determine which skill(s) to credit or blame, especially when a wrong answer is given. However, the special structure of the ASSISTment System gives teachers that very information. Since original questions are always followed by scaffolding questions, each addressing a single piece of the knowledge, when a student answered the original question wrong, we can rely on his responses to scaffolding questions to figure out exactly where the student has a misunderstanding and are able to track the specific knowledge component precisely. Moreover, students are using the ASSISTment System regularly in their normal math class during a school year, working on mathematics questions drawn from a pool of more than 1,400 questions generated from released state test items, local teachers, and project staff. The continuous usage allows us to collect more evidence of student's performance on every skill. Gierl et al. [24] proposed two directions for future research of cognitive assessment, of which one is to increase understanding of how to specify an appropriate grain size or level of analysis with a cognitive diagnostic assessment [32]. In this paper, we consider four skill models with different granularity, including a unidimensional model and a fine-grained model developed at WPI with 78 skills. The four models are structured with an increasing degree of specificity as the number of skills increases. The measure of model performance is the accuracy of the predicted MCAS test scores based on the assessed skills.

Given that the fine-grained model is composed of 78 skills, people might think that the model would naturally fit the data better than the skill models that contain far less skills,



5 skills your students doing well

WPI-5	WPI-39	WPI-78	Correct Rate		
Number Sense	N.10.8-computing-numbers	Addition		86%	392
	N.1.8-number-representations	Integers		85%	107
		Ordering-Numbers		79%	312
		Rounding		79%	164
	N.10.8-computing-numbers	Subtraction		76%	715

5 skills your students need more practice

WPI-5	WPI-39	WPI-78	Correct Rate		
Measurement	M.2.8-concerting-measures	Unit-Conversion		44%	1063
Algebra	A.7.8-solving-equations	Inequality-Solving		43%	593
Geometry	G.1.8-polygon-geometry	Sum-Of-Interior-Angles		42%	124
	G.4.8-pethagorean-theorem	Pythagorean-Theorem		37%	890
Data Analysis	D.2.8-data-representation-tech.	Venn-Diagram		37%	424

Fig. 3. A skill report showing skills students performed well or poorly.

maybe even overfit the data with so many free parameters. However, we were evaluating the effectiveness of the skill models over a totally different data from MCAS tests, namely, the external state tests as the testing set. Predicting students' scores on this test will be our gauge of model performance. Hence, we argue that overfitting would not be a problem in our approach.

1.3 Literature Review

Modeling student response data from intelligent tutoring systems has a long history (e.g., [12], [15]). Corbett and his colleagues employed a very detailed model of skills, but their system did not have questions tagged with more than one production rule [2]. Our collaborators [5] were engaged in trying to allow multimapping[4] using a version of the fine-grained model but reported their Linear Logistic Test Model (LLTM) does not fit well. Different from our approach, the model they applied does not track student performance over time. Almond et al. [1] *examine the application of Bayesian networks to Item Response Theory-based cognitive diagnostic modeling*. Bayesian networks have also been used to investigate the results of skill hierarchies using real-world data in intelligent tutoring systems (e.g., [23]) and simulated users (e.g., [10], [14]). Others (e.g., [8]), in the psychometrics field, have developed multidimensional Item Response Theory (IRT) models but these models do not allow multimapping.

There has also been a large interest in building cognitive diagnostic models. What we refer to as a "skill model" is referred to as "Q-Matrix" by some Artificial Intelligence researchers [6] and psychometricians [43], Croteau et al. (2004) called it "transfer model," while Cen et al. [9] and Gierl et al. [24] used the term "cognitive model." In all cases, a skill model is a matrix that relates questions to the skills needed to solve the problem. Such a model provides an interpretative framework to guide test development and psychometric analyses, so test performance can be linked to specific cognitive inferences about the examinees. Researchers in machine learning area have been using automatic/semiautomatic techniques to search for skill models, including the rule space method [43], the Q-matrix method [6], and Learning Factor Analysis (Cen et al. [9]). Though it addressed the same problem, our work is different in that we hand-coded the skill models and built the connection between skills and questions. This is similar to what Ferguson et al. [23] did in their work as they also associated problems with skills by hand, but they employed a different methodology.

Though different approaches have been adopted to develop skill models and model students' responses, as far as we know, little effort has been put in comparing different grain-sized skill models in the intelligent tutoring system area. The only work we are aware of that shows that by building fine-grained skill models researchers could build better fitting models was by Yun et al. [46] who developed an alternative curriculum framework. Their results of confirmatory factor analysis showed that the alternative framework fits data better suggesting the state's learning standards are subject to improvement.

Mislevy [30] described six steps in model-based reasoning in science. These steps, including model formation, elaboration, use, evaluation, revisions, and model-based inquiry, provide a framework for considering our progress in developing and refining cognitive models. Following these steps, the rest of the paper is organized as follows: In Section 2, we describe how the fine-grained model was developed and how it is currently being used in ASSISTment system. In Section 3, we evaluate the models by answering two research questions. Finally, we conclude in Section 4 and bring up the issue of model refinement and model-based inquiry as part of our future work.

2 Toward a Fine-Grained Skill Model

2.1 Developing a Fine-Grained Model for Eighth Grade MCAS

In April 2005, we staged a 7-hour-long "coding session," where we invited our subject matter expert and one of the authors to create a set of skills and use those skills to tag all of

4. A "multimapping" skill model, in contrast to a "single-mapping" model, allows one item to be tagged with more than one skill.

TABLE 1
Hierarchical Relationship among Skill Models

WPI-78	WPI-39	WPI-5	WPI-1
Inequality-solving	Setting-up-and-solving-equations	Patterns, Relations, and Algebra	Math (Unidimensional assessment)
Equation-solving			
Equation-concept			
…	…		
Plot-graph	Modeling-covariation		
X-Y-graph	Understand-line-slope-concept		
…	…	…	
Congruence	Understand-and-applying-congruence-and-similarity	Geometry	
Similar-triangles			
…	…		
Perimeter	Using-measurement-formulas-and-techniques	Measurement	
Area			
…	…		

the existing eighth grade MCAS items. 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 the limit that no one item would be tagged with more than three skills. It is not coincidence that many of our ASSISTment System questions have three scaffolding questions; we wanted the fine grainedness of the modeling to match the fine grainedness of the scaffolding. We wanted the scaffolds to have identifiability, meaning that each scaffolding question should be tagged with only one skill. Identifiability is important because when a student got a question tagged with two skills wrong, we will have a hard time coming up with a method that would be able to blame the lack of a single skill. As a matter of fact, in the data sources that we talk about in Section 3.1, the average number of skills tagged to main questions is 1.44 and the number is 1.03 for scaffolding questions, which means that for many questions in ASSISTment System, there is only one skill tagged with the main question. In such cases, each of the scaffolds was also tagged with that skill. There clearly is something a bit odd there, as an individual scaffolding question should be easier than the main item. We note this, but our modeling effort does not correct for the presumably wrong assumption that questions tagged with the same set of skills have the same degree of difficulty.

During the "coding session," the subject matter expert was free to create whatever skills she thought appropriate. She reviewed the items, solved the problems, and conducted a cognitive task analysis to identify what knowledge was needed to perform each task. When the coding session was over, we wound up with about a model of 106 skills. Yet, since only 78 out of the 106 skills were involved in the data used by this work, so we call this model the WPI-78. To create the coarse-grained models, we used the fine-grained model to guide us. We decided to use the same broad strands that are used by both the National Council of Teachers of Mathematics and the Massachusetts Department of Education. These strands are named:

1. "Patterns, Relations, and Algebra."
2. "Geometry."
3. "Data Analysis, Statistics, and Probability."
4. "Number Sense and Operations."
5. "Measurement."

The Massachusetts Department of Education actually tags each item with exactly one of the five strands, but our mapping was inferred from the WPI-78, so it was not the same as the state's mapping. Therefore, it was named WPI-5. Furthermore, we allowed multimapping, i.e., allowing an item to be tagged with more than one skill. An interesting piece of future work would be to compare our fit with the classification that the state uses. Similarly, we adopted the name of the 39 learning standards (nested inside the five strands) in the Massachusetts Curriculum Framework, associated each skill in WPI-78 to one of the learning standards, thus creating model WPI-39. After the students had taken the state tests, the state released the items in that test, and we had our subject matter expert tag up these items in WPI-5, WPI-39, and WPI-78.

The WPI-1, WPI-5, and WPI-39 models were derived from the WPI-78 model by nesting a group of fine-grained skills into a single category. Table 1 shows the hierarchical nature of the relationship among WPI-78, WPI-39, WPI-5, and WPI-1. The first column lists 9 of the 78 skills in the WPI-78 model.

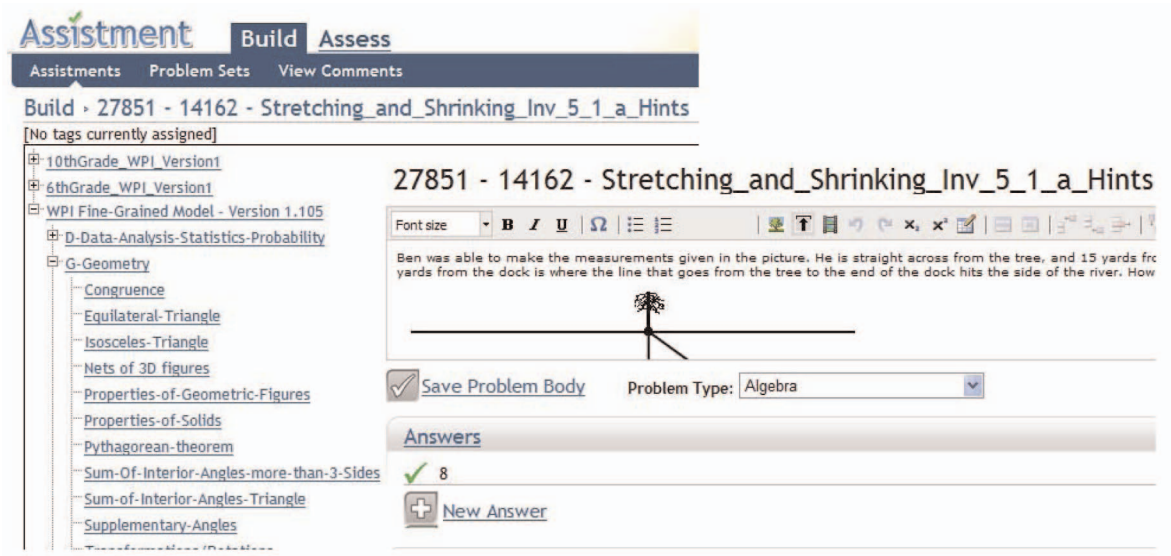


Fig. 4. A screen shot showing an item and a list of skills that could be tagged to the item in the builder.

For instance, *equation-solving* is associated with problems involving setting up an equation and solving it, while *equation-concept* is related to problems that have to do with equations in which students do not actually have to solve them. In the second column, we see how the two skills in WPI-78 are nested inside of "Patterns, Relations, and Algebra," which itself is one piece of the five skills that comprise the WPI-5 skill model.

As shown in Fig. 1, in the WPI-78 skill model, the first scaffolding question of item 19 from the 2003 MCAS test was tagged with "congruence," the second tagged with "perimeter," and the third tagged with "equation-solving." And the original question was associated with all three skills. When a student answered the original question wrong, we will examine his response to the scaffoldings to determine which skill is causing difficulty. In the WPI-39 model, the corresponding skills tagged are "Setting-up-and-solving-equations," "Understanding-and-applying-congruence-and-similarity," and "Using-measurement-formulas-and-techniques." In the WPI-5, the questions were tagged correspondingly with "Geometry," "Measurement," and "Patterns, Relations, and Algebra," and just one skill of "math" at the WPI-1.

2.2 Tagging Skills to Problems in the ASSISTment System

The ASSISTment Builder [26] provides technology support for authors to tag skills for the ASSISTment System question they build. This tool, shown in Fig. 4, provides a means to link certain skills to problems and to specify that solving the problem requires knowledge on that skill. The skills are organized in a hierarchical structure. The authors are allowed to browse the skills within each model and map the ones they select to a problem.

3 Statistical Analysis of the Skill Model

The fine-grained model gives teachers information about specific questions that are causing students trouble. But how did the finer grained model do on performance prediction? We address the question by analyzing log data using statistical methods.

3.1 Data Source

We collected data from 447[5] students, who used our system from 17 September 2004 to 16 May 2005 for, on average, 7.3 days (one period per day).[6] All these students worked on the system for at least six days (one session per day). We excluded data from the students' first day of using the system because they were learning how to use the system at that time. The item-level state test report was available for all these 447 students so that we were able to construct our predictive models on these students' data and evaluate the accuracy on state test score prediction. The original data set, corresponding to students' raw performance (before applying any "credit-and-blame" strategies as described below and not inflated due to the encoding used for different skill models), contained about 138,000 data points, among which around 43,000 come from original questions. On average, each student answered 87 MCAS (original) questions. We will refer to this data set as DATA-2005.

We obtained a similar data set for the usage of the ASSISTment System during the 2005-2006 as well. The data set involved 474 students who, on average, worked in the system for 5.5 days and answered 51 original questions. The item-level response data from the 2006 MCAS tests were available for these students too. This data set will be referred to as DATA-2006.

Both of the data sets are organized, so there can be one or multiple rows for every student response to each single question depending on which skill model we are interested in and how many skills the question is "tagged" with in that particular skill model. For instance, suppose a question is tagged with two skills in a model, then for each response made to the question, there would be two rows in the data set, with skill names listed in a separate column. Students'

5. The amount of data is limited by the maximum memory allowed by the open-source statistical package we used.

6. Given the fact that the state test was given on 17 May 2005, it would be inappropriate to use data after that day for the purpose of predicting state scores. Therefore, that data were not included in our data set.

TABLE 2
Sample Raw Data

Row ID	Student ID	State Test ID	question ID	WPI-78 skills	Original?	Response	Month Elapsed
1	950	2003- item 19	326	Congruence	Yes	0	1.32
2	950	2003- item 19	326	Perimeter	Yes	0	1.32
3	950	2003- item 19	326	Equation-Solving	Yes	0	1.32
4	950	2003- item 19	327	Congruence	No	0	1.32
5	950	2003- item 19	328	Perimeter	No	1	1.32
6	950	2003- item 19	329	Equation-Solving	No	0	1.32
7	950	2003- item 19	330	Equation-Solving	No	0	1.32
8	950	1999- item 27	1183	Perimeter	Yes	0	2.94
9	950	1999- item 27	1183	Area	Yes	0	2.94
10	950	1999- item 27	1184	Perimeter	No	1	2.94
11	950	1999- item 27	1185	Area	No	1	2.94

exact answers are not included. Instead, we use a binary column to represent whether the student answered the specified item correctly or not. No matter what the input type of the item is (multiple-choice or text-input), a "1" indicates a correct response, while a "0" means a wrong one. Additionally, a column is associated with each response, indicating the number of months elapsed since 17 September 2004 (or 1 September 2005) till the time when the response was made. Thus, in DATA-2005, the number of months elapsed for a response made on 17th September will be zero, and the number will be 1 for a response made at 17 October 2004, and so on. This gives us a longitudinal, binary response data set across the school year.

Table 2 displays 11 rows of the raw data for one student (system student ID $= 950$) who finished the item 19 (from 2003 state test, shown in Fig. 1) and item 27 (from 1,999 state test) on two different days. The first seven rows represent the student' responses to item 19 (with system ID for the original question being 326) and the remaining four rows show his responses to item 27 (with system ID for the original question being 1,183). We can see that since the original question of item 19 was tagged with three skills, "Congruence," "Perimeter," and "Equation-Solving," the student's response was duplicated in rows 1-3. Likewise, the original question of item 27 is tagged with two skills as shown in row 8 and row 9. If a student gives correct answer to an original question, the response column of all rows corresponding to that original question is marked as "1." If the student answered the original question wrong, he gets "0" in all the corresponding rows, and we will use his response to separate scaffolding questions to determine which skill to blame (assuming that each scaffolding question is tagged with only one skill). In the example, the student answered both the original questions incorrectly, thus got "0" in the response column of rows 1-3 and rows 8-9 and was presented with the scaffolding questions. The student did not do very well on the first item. He only gave a correct answer to the second scaffolding question (indicated by "1" in the response column of row 5) and failed on all others. In contrast, although the student did not get item 27 right on the first shot on the original question, the student wound up answering both scaffolding questions correctly. The WPI-78 is the skill model used here.

3.2 The Statistical Model Fitted to ASSISTments Data—Mixed-Effects Logistic Regression Model

We fit the mixed-effects logistic regression model, a longitudinal model, on the data to obtain an estimate of student knowledge on individual skills at a certain time, and for simplification, assuming that student knowledge was changing linearly over time. The mixed-effects logistic regression model is also referred to as the Generalized Linear Mixed-effects Model (GLMM) in the statistics literature and hierarchical models in Bayesian settings. It consists of both *fixed effects*, parameters corresponding to an entire population or repeatable levels of factors, and *random effects*, parameters corresponding to individual subject drawn randomly from a population. For dichotomous (binary in our case) response data, several approaches have been developed. These approaches use either a logistic regression model or a probit regression model and various methods for incorporating and estimating the influence of the random effects on individuals. Since we want to track individual student's development of skills over time and make predictions, we chose mixed-effects logistic regression model because it takes into account the fact that responses of one student to multiple items are correlated; moreover, the random effects allow the model to learn parameters for individual students separately. Hedeker and Gibbons [25] described 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.

As a statistical foundation of the mixed-effects generalization of the logistic regression model, we first present the simpler fixed-effects logistic regression model. Let p_i represent the probability of a positive response on an item for the ith individual. The probability of a negative outcome is then $1 - p_i$. Let $x_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$ denote the set of covariates and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ be the vector of corresponding regression coefficients. Then, the logistic regression model can be written as:

$$p_i = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}.$$

The model can also be represented in terms of log odds or *logit* of the probabilities, namely:

$$\log\left[\frac{p_i}{1-p_i}\right] = x_i'\beta.$$

In logistic regressions, the *logit* is called the link function because it maps the (0, 1) range of probabilities onto $(-\infty, +\infty)$ range of linear predictors. And by doing this, now the logistic regression model is linear in terms of the logit, though not in terms of the probabilities.

The simple logistic regression model can be generalized to be the mixed-effects model by introducing the random effects. Suppose TIME is the only covariate we care about in the model (*skill* can be introduced as a factor in the model in a similar way). The two-level representation of the model in terms of *logit* can be written as:

Level-1 (or within-person) model:

$$\log\left[\frac{p_{ij}}{1-p_{ij}}\right] = b_{0i} + b_{1i}{}^{*}TIME_{ij},$$

Level-2 (or between-person) model:

$$b_{0i} = \beta_0 + v_{0i},$$
$$b_{1i} = \beta_1 + v_{1i},$$

where p_{ij} is the probability that student i gives a correct answer at the jth opportunity of answering a question.

$TIME_{ij}$ refers the jth opportunity when student i answered a question. In our data, it is a continuous value representing the number of months (assuming 30 days in a month) elapsed since student i took his first action in the system.

Here, b_{0i}, b_{1i} denote the two learning parameters for student i. b_{0i} represents the "intercept" or how good is the student's initial knowledge; b_{1i} represents the "slope" which describes the change (i.e., learning) rate of student i.

β_0, β_1 are the fixed effects and represent the "intercept" and "slope" of the whole population average change trajectory.

v_{0i}, v_{1i} are the random effects and represent the student-specific variance from the population mean.

Such a model is often referred to as a "longitudinal model" [41] since TIME is introduced as a predictor of the response variable, which allows us to investigate change over time. The models were fitted in R [36] using *lmer()* function in *lme4* package [7] and "logit" was used as the link function. In this model, we introduced skills as fixed-effect factor and TIME (*monthElapsed*) as both a fixed effect and a random effect in order to learn both the learning rate per month for the whole group of students, on average, and the variation of each individual student. We also included the interaction between *skills* and *monthElapsed*, which told the model to learn students' average learning rate separately for each skill. Notice that we did not include skills as random effect, which meant the model assumed that a student's learning rate did not vary over different skills.[7]

After the model was constructed, the fixed effects for the whole group (i.e., β_0, β_1 in the above two-level model), including an intercept, a coefficient for the *monthElapsed* covariate, four coefficients for the *skills*, one for each skill in the WPI-5 model, and four coefficients for the interaction term, and the random effects for each student (i.e., v_{0i}, v_{1i}), including an intercept indicating a student's incoming knowledge and a slope (coefficient for *monthElapsed* as a random effect) indicating the student's overall learning rate per month, were extracted. Then, the two learning parameters "intercept" and "slope" (i.e., b_{0i} and b_{1i} in the model above) were calculated for each individual student and each skill. Given this, we can apply the model on the items in the state test to estimate students' response to each of them.

3.3 Predicting State Test Scores

After the model is fit, we have skill levels of the students based on their online ASSISTment data using the specified skill model. We then apply the model on the actual state test. All the items in the state tests have been tagged in all of the four skill models by our subject matter expert.[8] To predict a student's test score when a particular skill model is adopted, we will first find the fractional score the student can get on each individual item, and then, sum the "item-score" up to acquire a total score for the test. So, how did we predict their state test item score?

Given a student's learning parameters on all skills and the exact test date of MCAS, we can calculate the probability of positive response from the student to an item tagged with any single skill. In the case that an item was tagged with more than one skill, we picked the skill that gave us the lowest probability among all the skills that apply to the item[9] for that student (the hardest skill for the student). Thus, we obtained the probability of positive response to any particular item in the state test. In our approach, a student's probability of correct response for an item was used directly as the fractional score to be awarded on that item for the student. We summed item scores up to produce the total points awarded on the test. For example, if the probability of an item marked with Geometry is 0.6, then 0.6 points were added to the sum to produce the points awarded. This sum of these points was what we use as the predicted state test score.[10]

The prediction function we built using the existing data are also intended to work well in future years, and so for reasons of interpretability, the prediction error function chosen was mean absolute deviation (MAD). This measure was suggested by Junker, a statistician from Carnegie Mellon University [27]:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|MCAS_i - prediction_i|,$$

7. This is just a simplifying assumption. Of course, in reality, it is possible that a student might learn one skill (e.g., *perimeter*) faster than another one (e.g., congruence).

8. All the tagging was done after the MCAS items were released without any reference to the modeling process described in this paper.

9. We admit that there are other approaches dealing with multimapped items. For instance, using Bayesian Networks is a reasonable way to deal with this situation. Pardos et al. [34] use this approach and got similar results that fine-grained models enable better predictive models.

10. We think that it might be useful to discuss our model from a more qualitative point of view. Is it the case that if you tag an item with more skills, does that mean our model would predict that the item is harder? The answer is no, in the sense that if you tagged a bunch of items with an easy skill (i.e., one easier than what the item was currently tagged with), which would not change our model's prediction at all. This makes qualitative sense, in that we believe the probability of getting a question correct, is given by the probability of getting correct the most difficult skill associated with that question.

where *MCASi* is the actual MCAS score of the ith student, and *predictioni* is the predicted score from the prediction function being evaluated. For every model, we subtracted each student's real test score from his predicted score, took the absolute value of the difference, and averaged them to get the MAD. We also calculated a normalized metric named % Error by dividing the MAD by the full score:

$$\%Error = MAD/(MaxRawScore),$$

where "*MaxRawScore*" is the maximum raw score possible with the MCAS questions used. The MCAS state test consists of five open response, four short answer, and 30 multiple choice questions. The max score is 54 points if all 39 MCAS questions are considered, since some are scored wrong/right and some are scored with partial credit. In our case, only the multiple-choice and short-answer questions are used with regard to the fact that currently, open response questions are not supported in our system. This makes a full score of 34 points with one point earned for a correct response on an item. For the students in our 2005 data set, the mean score out of 34 points was 17.9 (standard deviation $= 7.1$). For the students in 2006 data set, the mean score was 18.8 (standard deviation $= 7.8$).

3.4 Research Question 1: Does Adding Scaffolding Questions Help?

Research Question 1 (**RQ1**). Would adding response data to scaffolding questions help us to do a better job of tracking students' knowledge and more accurately predicting state test scores, compared to only using the original questions? Because the scaffolding questions break the test question down into a series of simpler tasks that directly assess fewer knowledge components, we believe that the ASSISTment System can do a more accurate assessing job. This hierarchal breakdown of knowledge provides a much finer grained analysis than is currently available. We think that getting an answer to RQ1 would help us properly evaluate the second and more important research question described in Section 3.5.

3.4.1 Scaffolding Credit and Partial Blame

We started our work examining only students' responses to original questions. And then, we brought up RQ1, asking ourselves if we can improve our models by including students' response to the scaffolding questions. As discussed in Section 1, adding in scaffolding responses creates a good chance for us to detect exactly which skills are the real obstacles that prevent students from correctly answering the original questions. This would be especially useful when we utilize a finer grained model.

Since the scaffolding questions show up only if the students answer the original question incorrectly, their responses to the scaffolding questions are explicitly logged. However, if a student gets an original question correct, he is only credited for that one question in the raw data. To deal with the "selection effect," we introduced the compensation strategy of "scaffolding-credit": Scaffolding questions are also marked correct if the student gets the original questions correct.

An important thing we need to determine when using a multimapping model (in which one item is allowed to be tagged with more than one skill) is which skills to blame when a student answered an item tagged with multiple skills incorrectly. Intuitively, the tutor may want to blame all the skills involved; however, this would be unfair to those relatively easy skills when they are tagged to some compound, hard items. To avoid this problem, we applied the "partial blame" strategy: If a student got such an item wrong, the skills in that item will be sorted according to the overall performance of that student on those skills and only the skill on which that particular student showed the worst performance will be blamed.

When evaluating a student's skill levels, both original questions and scaffold responses are used in an equal manner and they have the same weight in evaluation.

3.4.2 Results

Recall that RQ1 asked whether adding response data to scaffolding questions can help us to do a better job of tracking students' knowledge and more accurately predicting state test scores. To answer RQ1, we first trained mixed-effects logistic regression models using the data set that only includes original questions response; one regression model for each skill model. Then, we replicated the training process but used the data set that was constructed by including responses to scaffolding questions and applying the "credit-and-blame" strategy described as above. Again, models were trained for all three skill models.

It turns out that better fitted models, as measured by % Error, on the state test can always be obtained by using scaffolding questions. In particular, when using the WPI-1 on DATA-2005, the mean decrease of % Error is 1.91 percent after scaffolding questions were introduced; for WPI-5, the decrease is 1.21 percent; and the decrease of % Error is 2.88 percent for the WPI-39; and 5.79 percent for the WPI-78 which is the biggest improvement. We then did paired t-tests between the % Error terms for the 447 students and found that the improvements are statistically significant in all the four cases as summarized in Table 3. We noticed the same effect in DATA-2006. As shown in Table 3, the improvement on % Error is statistically reliable on all of the four models. (Please read across the columns for an answer to RQ1. Reading across the rows is the answer to RQ2 that we will describe in the next section.)

This drop-down of % Error (also MAD) makes sense for two reasons. One is that by using the response data to scaffolding questions, we are using more of the data we collected. A second reason is that the scaffolding questions help us to do a better job of dealing with credit-and-blame problems. To get more "identifiability" per skill, in the next section, we use the "full" response data (with scaffolding question responses added in) to try to answer the question of whether finer grained models predict better.

Sharp readers may have noticed that the MAD of WPI-39 model for DATA-2006 is lower than that of WPI-78, yet % Error of the WPI-39 model is higher than % Error of the WPI-78 model. This is because the two multiple-choice items in 2006 MCAS test, item 13 and item 26, were tagged with the skills "N.6.8-understanding-absolute-value" and "P.9.8-modeling-covariation," respectively, yet, none of the ASSISTment System items were tagged by the same two skills, which means that we do not have training data to track student knowledge on the two skills. Therefore, we ignored the two items when predicting students' total score of 2006 MCAS test using the WPI-39 model. This reduces the total number of MCAS items of the WPI-39 to 32. The % Error of

TABLE 3
The Effect of Using Scaffolding Questions on DATA-2005 and DATA-2006

Skill Model	MAD		% Error (MAD/#items)		Δ %Error	p-value of paired t-test
	Orig. Response	Orig.+ Scaffolding Response	Orig. Response	Orig.+ Scaffolding Response		
DATA-2005						
WPI-1	5.07	4.42	14.91%	13.00%	1.91%	0.008
WPI-5	4.78	4.37	14.06%	12.85%	1.21%	0.049
WPI-39	5.20	4.22	15.29%	12.41%	2.88%	<0.0001
WPI-78	6.08	4.11	17.75%	12.09%	5.79%	<0.0001
DATA-2006						
WPI-1	6.81	6.58	20.05%	19.37%	0.63%	0.001
WPI-5	6.76	6.51	19.88%	19.14%	0.74%	<0.0001
WPI-39	5.98	4.83	18.68%	15.10%	3.58%	<0.0001
WPI-78	5.58	4.99	16.91%	14.70%	2.21%	<0.0001

the WPI-39 model is calculated by MAD/32, while the % Error of the other models are calculated by MAD/34.

Does an error rate of 12.09 percent on the WPI-78 seem impressive or poor? What is a reasonable goal to shoot for? Zero percent error? For comparison reason, we created a baseline estimation of students' MCAS test scores by first computing students' overall percent correct on original questions, and multiplied the % correct with the full score. Under this "dumb" approach, the % Error was 17.26 percent for DATA-2005 and 21.47 percent for DATA-2006. In [15], we reported on a simple simulation on how well one MCAS test was at predicting another MCAS test. We did not have access to data for a group of students who took two different versions of the MCAS test to measure this, so we estimated it by taking students' item-level scores on MCAS, randomly splitting the 34 multiple-choice items in the test into two halves, and then, using their scores on the first half to predict the second half. This process was repeated five times, and, on average, the % Error was 11 percent, suggesting that a 12 percent error rate is looking somewhat impressive.

3.5 Research Question 2: Does the Finer Grained Model Predict Better?

Research Question 2 (**RQ2**). How does the finer grained skill model (WPI-78) do on estimating external test scores compared to the other skill models?

We think that an answer to RQ2 that says that a finer grained model allows for better modeling/prediction would have important societal implications (e.g., regarding tracking student performance and reporting to teachers).

3.5.1 Does WPI-78 Fit Better than the Coarser-Grained Models?

To answer RQ2, we compared the four mixed-effects regression models (trained on the "full" data set with scaffolding questions used) fitted using the four different skill models. As shown in Table 4, the WPI-78 had the best result, followed by the WPI-39, WPI-5, and WPI-1. % Error dropped down when a finer grained model was used from WPI-1 to WPI-5, and then, from WPI-39 to WPI-78.

To see if the % Error was statistically significantly different for the models, we compared each model with every other model. We did paired t-tests between the % Error terms for the 447 students in DATA-2005 and also the 474 students in DATA-2006. We found out that in DATA-2005, the WPI-78 did as well as the WPI-39 ($p = 0.21$), and they both predicted MCAS score reliably better than the WPI-5 and WPI-1. In DATA-2006, the WPI-78 model is statistically reliably better than the WPI-39, WPI-5, and WPI-1 ($p < 0.001$ in all cases), and WPI-1 is statistically reliably worse on predicting MCAS scores than the other models ($p < 0.0001$). This suggested that finer grained skill models were helpful in tracking students' knowledge over time.

We want to address that our results on student performance prediction are by no means the best. As a matter of fact, we trained an Item Response Theory [42] model that has been widely used in traditional testing area by psychometricians as a control. We fit the simplest model, the Rasch model that models student i dichotomous response ($0 = \text{wrong}, 1 = \text{correct}$) to problem j as a logistic function of the difference between student proficiency (θ_i) and problem difficulty (β_j), on our online data. The fitted model gave us an estimate of math proficiency for every individual student which allows us to compute the predicted MCAS score assuming that every item in MCAS has an average difficulty ($\beta = 0$). In Table 4, *IRT-2005* refers to the IRT modeling condition for DATA-2005, and *IRT-2006* refers to the IRT modeling for DATA-2006. As we can see, the % Error of the Rasch model for DATA-2005 is 12.82 percent, marginally higher than that of the WPI-78, 12.09 percent ($p = 0.10$). Yet, the Rasch model

TABLE 4
Evaluating the Accuracy of Skill Models

Skill Model	MAD	95% Confidence Interval for MAD	% Error	
DATA-2005				
WPI-1	4.42	[4.12, 4.72]	13.00%	p=0.006
WPI-5	4.37	[4.07, 4.66]	12.85%	p<0.0001
WPI-39	4.22	[3.94, 4.50]	12.41%	p=0.21
WPI-78	4.11	[3.84, 4.39]	12.09%	p=0.10
IRT-2005	4.36	[4.04, 4.68]	12.82%	
DATA-2006				
WPI-1	6.58	[6.18, 6.99]	19.37%	p<0.0001
WPI-5	6.51	[6.11, 6.90]	19.14%	p<0.0001
WPI-39	4.83	[4.56, 5.11]	15.10%	p=0.0001
WPI-78	4.99	[4.71, 5.28]	14.70%	p=0.03
IRT-2006	4.67	[4.34, 4.99]	13.7%	

did better in the next year where the % Error (13.70 percent) is reliably higher ($p = 0.03$) than that of the WPI-78 (14.70 percent). Other than the IRT model, we have also contrasted our result on DATA-2005 with the result produced by Bayesian network approach that dealt with skills associated with one item conjunctively using "AND" gate [34]. The "AND" gate signifies that all the skills must be known in order for the questions to be answered correctly. Pardos et al. [35] confirm that the "conjunctive" hypothesis. During the comparison process, we found out that our approach did better than the Bayesian networks approach when the WPI-1 and WPI-5 models were used, and the two approaches are comparable when the WPI-39 and WPI-78 were used. Specifically, for the WPI-39 model, % Error of the Bayes approach is 12.05 percent, lower than what we got (12.41 percent); yet for the WPI-106 model, % Error of the Bayes approach is 13.75 percent, higher than our result of 12.09 percent.

As a measure of internal fit, we calculated the average absolute residual for each model fitted on the data. For data of both years, the WPI-78 fits best. Since the WPI-78 model contains far more skills than other models, one might think the model won simply because of the large number of parameters. Therefore, as a sanity check, we generated a Random-WPI-78 model in which items are randomly mapped with skills from the WPI-78 model. It turned out that the random model did reliably worse than the WPI-78 model (and also the WPI-39), both in MCAS score prediction and the internal fit.[11] Readers may have noticed in Table 3 that when only response data on original questions were used, the order changed for DATA-2005: The WPI-5 still did better than WPI-1. However, the prediction error gets worse when the WPI-39 or WPI-78 models were used. Our interpretation of this is that when only original responses were used, individual skills do not get as much identifiability; it only makes sense to make fine-grained skill models, if you have questions that can be tagged with just a single skill. Another reason why finer grained models might not fit the data as well would be the fact that the finer grained model has fewer data points per skill, so there is a trade-off between the number of skills you would like and the precision in the estimates.

Comparing the results that we got using DATA-2005 and those using DATA-2006, we noticed two things changed. First, the order of prediction accuracy differs when only original questions were used. The finer grained models still track student knowledge better than coarser-grained models when DATA-2006 was used; yet it is not the case when DATA-2005 was used. Second, the prediction error was much higher in the year 2005-2006 than in the previous year. Third, the effectiveness of the IRT model reduced in the year 2006. One possible reason is that we have fewer training data points for each student in the year 2005-2006 (5.5 sessions and 51 problems done versus 7.3 sessions and 87 problems done). Additionally, the problem sets administered to students in the two years are not the same.

3.5.2 How Well Does the Model WPI-78 Fit the Data?

When using logistic regression, the statistical packages allow the user to analyze which of the parameters seem to have good fitting values. We now turn to do a little more analysis on the WPI-78 to see how good our model is. In our model, each skill gets one coefficient indicating the skill's "intercept" and one for the skill's "slope." The first of these, the intercept, allows us to model that some skills start the

11. It is common to report the value of a model by using a metric that balances model fit and model complexity such as Bayesian Information Criterion (BIC). For instance, Cen et al. [9] and Ferguson et al. [23] both used BIC to compare different models. However, because the size of the data sets was different when we used the different models; the finer grained models add additional rows for all questions that are tagged with more than one skill, while BIC only makes sense when the data are meant to be the exact same size. For the same reason, we did not conduct ANOVA on the results.

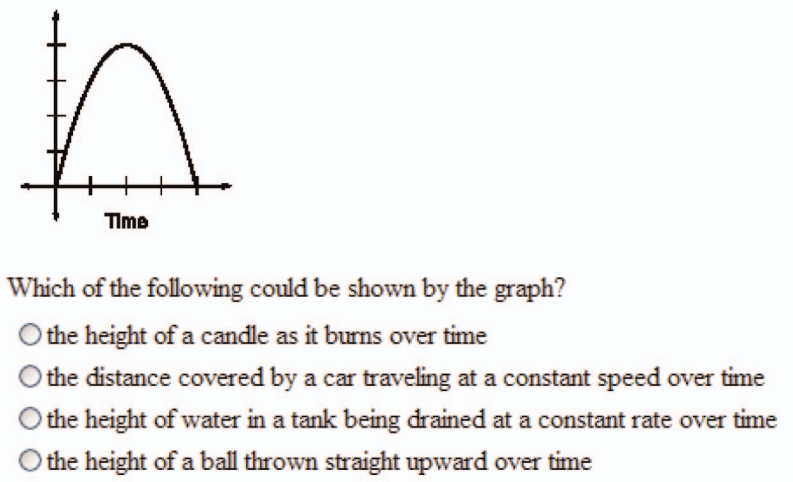


Fig. 5. A question tagged with the skill "Qualitative-Graph-Interpretation."

year with students knowing them better, while the slope allows for the fact that some skills are learned more quickly than others. Our model shows that for students who used the system in the school year 2004-2005, the easiest skills are "Subtraction," "Division," and "Simple-Calculation," while the skill that had the hardest incoming difficulty was "Qualitative-Graph-Interpretation" (as shown in Fig. 5). We also looked at the fits on the slopes for each skill. The skill that showed the steepest rate of learning during the course of the year was "Sum-of-Interior-Angles-Triangle" (e.g., "what is the sum of the angles inside of a triangle?"). It seems quite plausible that students learned a good amount related to this skill as we noticed in a classroom a poster that said "The sum of the interior angles in a triangle is 180" clearly indicating that this was a skill that teachers were focused on teaching. The skill that showed the least learning was called " Equation-Concept" (as shown in Fig. 6). Out of the 78 skills, seven coefficients predicted "unlearning" (i.e., the slopes are negative), which presumably raised a sign of overfitting, or that the tagging of the skills in the skill model was not quite right. In the future, we will investigate automating the process to remove such skills from the model and refit the data.

Considering the accuracy of fit, we noticed that the model obtained a high accuracy on predicting student response on items tagged with the simple skills (e.g., Division, Subtraction), yet not so good at tracking student knowledge on skills "Of-Means-Multiply," "Interpreting-Linear-Equations," or "Inequality-Solving." We speculated that skills that had less data for them would be more likely to be poorly fit. We did a correlation to see if the skills that were poorly fit were the same skills that had a relatively smaller numbers of items, but surprisingly the correlation was very weak. Other reasons that a skill might have a poorly fit slope would be that we tagged items with the same skill names that share some superficial similarity, but do not have the same learning rates. This analysis suggests some future work in refining the WPI-78 model; for instance, one possible refinement is to merge "equation-concept" with "equation-solving" (i.e., delete the "equation-concept" skill from the model and map all items tagged with "equation-concept" to "equation-solving"). Computational techniques such as Learning Factors Analysis [9] provide a way to manipulate the skill model, thus substantially improve the model fit to data.

All in all, we make no claim that the fine-grained model we created represented the best fitting model possible. Nevertheless, we stand by the claim that this model, taken in total, is good enough that it can produce good fit to the data and make good predictions of the MCAS scores, indicating that the model is useful, even given the flaws that might exist in it.

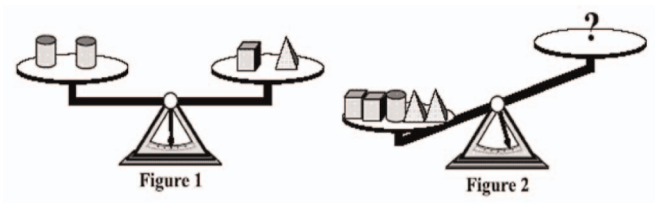


Fig. 6. A question tagged with the skill "Equation-Concept."

4 Conclusion and Future Work

In this paper, we presented evidence that we can over time track students' skills in an intelligent tutoring system like the ASSISTment System. We found pretty solid evidence that using students' responses to scaffolding questions was helpful in tracking students' knowledge (RQ1) and fine-grained models can better predict MCAS scores than coarser ones (RQ2). As discussed in the paper, teachers want reports by skills at fine-grained level, and this is the first evidence we have saying that our skill mappings are good enough to better predict a state test than some less fine-grained models.

This paper's main contribution is demonstrating the value of a fine-grained versus more coarse-grained models within intelligent tutoring systems. We rigorously evaluated the effect of granularity of the skill models and validated the usage of a fine-grained skill models. Another contribution of this paper is that we demonstrated how skill models in intelligent tutoring systems can be used to predict standardized test scores and we have shown that we do not do a

worse job of predicting students knowledge if we use a fine-grained model. This is important because teachers want finer grained information, but it is usually assumed that if you want to predict a large-scale state test based on a fine-grained model, you are not likely to succeed.

We believe that the ASSISTment System can be an even better predictor of state test scores because of this work. Yet, the more general implication from this research has to do with the possibility of building continuous learning and assessment systems. In [22], we reported that the ASSISTment System can be a better assessor after accounting for information such as the amount of assistance students required and their help-seeking behavior. The results presented in this paper further showed that not only can reliable assessment and instructional assistance be effectively blended in a tutoring system, but also, more importantly, such a system can provide teachers with useful fine-grained student-level knowledge they can reflect on and adjust their pedagogy. Recently, in an interview with US News & World Report [38], Secretary of Education Arne Duncan weighed in on the NCLB Act and called for continuous assessment. He mentioned that he is concerned about overtesting, and feels that fewer, better tests would be more effective. He wants to develop better data management systems that will help teachers track individual student progress in real time so that teachers and parents can assess and monitor student strengths and weaknesses. Our studies implies that it is possible for the states to develop such a system similar to the ASSISTment System that does all three of these things at the same time: 1) accurately assesses students; 2) gives fine-grained feedback that is more cognitively diagnostic; and 3) saves classroom instruction time by assessing students while they are getting tutoring.

Now that we are getting reliable results indicating the value of these models, we will seriously consider using these models in selecting the next best problem to present a student with. Existing literature has shown that creating an accurate model of a students' knowledge can be quite difficult due to various sources of uncertainty caused by factors such as multiple sources of student errors, varied problem solving strategies, problems associated with more than one skill, careless slip and lucky guesses, learning and forgetting [28], requiring the time of experts to create, and then test these models on students. The first model is the best guess and should be iteratively refined after usage in intelligent tutoring systems. The expert-built models are subject to the risk of "expert blind spot" [31]. We are happy to see that our first cognitive model fits well on student performance data. Nevertheless, we still feel that we can probably refine the fine-grained model to be more accurate. We have found that maintaining a cognitive model is difficult in a system where new questions are being added everyday. For future work, we plan to improve the model iteratively and use student performance data to evaluate the fitness of the models in each cycle, focusing on the less well-fitted skills.

ACKNOWLEDGMENTS

This research was made possible by the US Department of Education, Institute of Education Science, "Effective Mathematics Education Research" program grant #R305K03140 and "Making Longitudinal Web-based Assessments Give Cognitively Diagnostic Reports to Teachers, Parents, and Students while Employing Mastery learning" program grant #R305A070440, the US Office of Naval Research grant #N00014-03-1-0221, US National Science Foundation (NSF) CAREER award to Neil Heffernan, and the Spencer Foundation. All the opinions in this paper 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 and 2005-2006 WPI/CMU ASSISTment Teams.

REFERENCES

[1] R.G. Almond, L.V. DiBello, B. Moulder, and D. Zapata-Rivera, "Modeling Diagnostic Assessment with Bayesian Networks," *J. Educational Measurement,* vol. 44, no. 4, pp. 341-359, 2007.

[2] J.R. Anderson, *Rules of Mind.* Lawrence Erlbaum Assoc., 1993.

[3] J.R. Anderson and C. Lebiere, *The Atomic Components of Thought.* Lawrence Erlbaum Assoc., 1998.

[4] J.R. Anderson, A.T. Corbett, K.R. Koedinger, and R. Pelletier, "Cognitive Tutors: Lessons Learned," *J. Learning Sciences,* vol. 4, no. 2, pp. 167-207, 1995.

[5] E. Ayers and B.W. Junker, "Do Skills Combine Additively to Predict Task Difficulty in Eighth Grade Mathematics?" *Proc. Educational Data Mining: Papers from the Am. Assoc. for Artificial Intelligence (AAAI) Workshop,* J. Beck, E. Aimeur, and T. Barnes, eds., pp. 14-20, 2006.

[6] T. Barnes, "Q-Matrix Method: Mining Student Response Data for Knowledge," *Educational Data Mining: Papers from the 2005 Am. Assoc. for Artificial Intelligence (AAAI) Workshop,* J. Beck, ed., 2005.

[7] D. Bates, "Linear Mixed Model Implementation in lme4," Manuscript, Univ. of Wisconsin, May 2007.

[8] R.D. Bock, R. Gibbons, and E.J. Muraki, "Full Information Item Factor Analysis," *Applied Psychological Measurement,* vol. 12, pp. 261-280, 1988.

[9] H. Cen, K. Koedinger, and B. Junker, "Automating Cognitive Model Improvement by A*Search and Logistic Regression," *Proc. Educational Data Mining: Papers from the 2005 Am. Assoc. for Artificial Intelligence (AAAI) Workshop,* J. Beck, ed., 2005.

[10] J. Collins, J. Greer, and S. Huang, "Adaptive Assessment of Using Granularity Hierarchies and Bayesien Nets," *Proc. Int'l Conf. Intelligent Tutoring Systems,* pp. 569-577, 1996.

[11] J. Confrey, A. Valenzuela, and A. Ortiz, "Recommendation to the Texas State Board of Education on the Setting of TAKS Standards: A Call to Responsible Action," http://www.syrce.org/State_Board.htm, 2002.

[12] A.T. Corbett, J.R. Anderson, and A.T. O'Brien, "Student Modeling in the ACT Programming Tutor," *Cognitively Diagnostic Assessment,* P. Nichols, S. Chipman, and R. Brennan, eds., Lawrence Erlbaum Assoc., 1995.

[13] E. Croteau, N.T. Heffernan, and K.R. Koedinger, "Why Are Algebra Word Problems Difficult? Using Tutorial Log Files and the Power Law of Learning to Select the Best Fitting Cognitive Model," *Proc. Seventh Int'l Conf. Intelligent Tutoring Systems,* pp. 240-250, 2004.

[14] B. Daniel, D. Zapata-Rivera, and G. McCalla, "A Bayesian Belief Network Approach for Modeling Complex Domains," *Bayesian Network Technologies: Applications and Graphical Models,* A. Mittal and A. Kassim, eds., pp. 13-41, IRM Press, 2007.

[15] K.L. Draney, P. Pirolli, and M. Wilson, "A Measurement Model for a Complex Cognitive Skill," *Cognitively Diagnostic Assessment,* P. Nichols, S. Chipman, and R. Brennan, eds., Lawrence Erlbaum Assoc., 1995.

[16] S.E. Embretson, "Structured Rasch Models for Measuring Individual-Difference in Learning and Change," *Int'l J. Psychology,* vol. 27, nos. 3/4, pp. 372-372, 1992.

[17] M. Feng, N.T. Heffernan, and K.R. Koedinger, "Predicting State Test Scores Better with Intelligent Tutoring Systems: Developing Metrics to Measure Assistance Required," *Proc. Eighth Int'l Conf. Intelligent Tutoring Systems,* M. Ikeda, K.D. Ashley, and T.-W. Chan, eds., pp. 31-40, 2006.

[18] M. Feng, N.T. Heffernan, M. Murali, and C. Heffernan, "Using Mixed-Effects Modeling to Compare Different Grain-Sized Skill Models," *Proc. Educational Data Mining: Papers from the Am. Assoc. for Artificial Intelligence (AAAI) Workshop,* J. Beck, E. Aimeur, and T. Barnes, eds., pp. 57-66, 2006.

[19] M. Feng and N. Heffernan, "Towards Live Informing and Automatic Analyzing of Student Learning: Reporting in ASSISTment System," *J. Interactive Learning Research,* vol. 18, no. 2, pp. 207-230, 2007.
[20] M. Feng, J. Beck, N. Heffernan, and K. Koedinger, "Can an Intelligent Tutoring System Predict Math Proficiency as well as a Standardized Test?" *Proc. First Int'l Conf. Education Data Mining,* R.S.J.d. Baker and J.E. Beck, eds., pp. 107-116, 2008.
[21] M. Feng, N. Heffernan, J. Beck, and K. Koedinger, "Can We Predict Which Groups of Questions Students Will Learn From?" *Proc. First Int'l Conf. Education Data Mining,* R.S.J.d. Baker and J.E. Beck, eds., pp. 218-225, 2008.
[22] M. Feng, N.T. Heffernan, and K.R. Koedinger, "Addressing the Assessment Challenge in an Online System That Tutors as It Assesses," to be published in *User Modeling and User-Adapted Interaction: The J. Personalization Research.*
[23] K. Ferguson, I. Arroyo, S. Mahadevan, B. Woolf, and A. Barto, "Improving Intelligent Tutoring Systems: Using Expectation Maximization to Learn Student Skill Levels," *Proc. Eighth Int'l Conf. Intelligent Tutoring Systems,* M. Ikeda, K.D. Ashley, and T.-W. Chan, eds., pp. 453-462, 2006.
[24] M.J. Gierl, C. Wang, and J. Zhou, "Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees' Cognitive Skills in Algebra on the SAT," *J. Technology, Learning, and Assessment,* vol. 6, no. 6, http://www.jtla.org, 2008.
[25] D. Hedeker and R.D. Gibbons, *Longitudinal Data Analysis.* John Wiley & Sons, 2006.
[26] N.T. Heffernan, T.E. Turner, A.L.N. Lourenco, M.A. Macasek, G. Nuzzo-Jones, and K.R. Koedinger, "The ASSISTment Builder: Towards an Analysis of Cost Effectiveness of ITS Creation," *Proc. 19th Int'l Florida Artificial Intelligence Research Soc. Conf. (FLAIRS '06),* G. Sutcliffe and R. Goebel, eds., pp. 515-520, 2006.
[27] B. Junker, "Using On-Line Tutoring Records to Predict End-of-Year Exam Scores: Experience with the ASSISTments Project and MCAS Eighth Grade Mathematics," *Assessing and Modeling Cognitive Development in School,* R.W. Lissitz, ed., JAM Press, 2007.
[28] S. Katz, A. Lesgold, G. Eggan, and M. Gordin, "Modeling the Student in Sherlock II," *Int'l J. Artificial Intelligence in Education,* vol. 3, no. 4, pp. 495-518, 1992.
[29] M. Militello, S. Sireci, and J. Schweid, "Intent, Purpose, and Fit: An Examination of Formative Assessment Systems in School Districts," *Proc. Am. Educational Research Assoc. Ann. Meeting,* 2008.
[30] R. Mislevy, "Cognitive Psychology and Educational Assessment," *Educational Measurement,* fourth ed., R.L. Brennan, ed., Am. Council on Education, 2006.
[31] M.J. Nathan and K.R. Koedinger, "An Investigation of Teachers' Beliefs of Students' Algebra Development," *Cognition and Instruction,* vol. 18, no. 2, pp. 209-237, 2000.
[32] *Cognitively Diagnostic Assessment,* P.D. Nichols, S.F. Chipman, and R.L. Brennan, eds. Lawrence Erlbaum Assoc., 1995.
[33] L. Olson, "State Test Programs Mushroom as NCLB Mandate Kicks," *Education Week,* pp. 10-14, Nov. 2004.
[34] Z. Pardos, M. Feng, N.T. Heffernan, and C. Heffernan-Lindquist, "Analyzing Fine-Grained Skill Models Using Bayesian and Mixed Effect Methods," *Proc. 13th Conf. Artificial Intelligence in Education,* R. Luckin and K. Koedinger, eds., pp. 626-628, 2007.
[35] Z.A. Pardos, N.T. Heffernan, C. Ruiz, and J. Beck, "The Composition Effect: Conjunctive or Compensatory? An Analysis of Multi-Skill Math Questions in ITS," *Proc. First Int'l Conf. Educational Data Mining,* R.S.J.d. Baker and J.E. Beck, eds., 2008.
[36] R Development Core Team "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, http://www.r-project.org, 2007.
[37] A. Raftery, "Bayesian Model Selection in Social Research," *Sociological Methodology,* vol. 25, pp. 111-163, 1995.
[38] E. Ramírez and K. Clark, "What Arne Duncan Thinks of No Child Left Behind: The New Education Secretary Talks about the Controversial Law and Financial Aid Forms," http://www.usnews.com/articles/education/2009/02/05/what-arne-duncan-thinks-of-no-child-left-behind.html, 2009.
[39] L. Razzaq and N.T. Heffernan, "Scaffolding vs. Hints in the Assistment System," *Proc. Int'l Conf. Intelligent Tutoring Systems,* M. Ikeda, K.D. Ashley, and T.-W. Chan, eds., pp. 635-644, 2006.
[40] L. Razzaq, N.T. Heffernan, and R.W. Lindeman, "What Level of Tutor Interaction Is Best?," *Proc. 13th Conf. Artificial Intelligence in Education,* R. Luckin and K. Koedinger, eds., pp. 222-229, 2007.
[41] J.D. Singer and J.B. Willett, *Applied Longitudinal Data Analysis: Modeling Change and Occurrence.* Oxford Univ. Press, 2003.
[42] E.S. Tan, T. Imbos, and R.J.M. Does, "A Distribution-Free Approach to Comparing Growth of Knowledge," *J. Education Measurement,* vol. 31, no. 1, pp. 51-65, 1994.
[43] K.K. Tatsuoka, "Toward an Integration of Item Response Theory and Cognitive Error Diagnosis," *Diagnostic Monitoring of Skill and Knowledge Acquisition,* N. Frederiksen, R. Glaser, A. Lesgold, and M.G. Shafto, eds., pp. 453-488, Lawrence Erlbaum Assoc., 1990.
[44] *Handbook of Modern Item Response Theory,* W.J. Van Der Linden, and R.K. Hambleton eds. Springer Verlag, 1997.
[45] E.C. Wylie and J. Ciofalo, "Supporting Teachers' Use of Individual Diagnostic Itemsm," http://www.tcrecord.org/PrintContent.asp?ContentID=15363, 2008.
[46] J.T. Yun, J. Willet, and R. Murnane, "Accountability-Based Reforms and Instruction: Testing Curricular Alignment for Instruction Using the Massachusetts Comprehensive Assessment System," *Proc. Ann. Am. Educational Research Assoc. Meeting,* 2004.

Mingyu Feng received the BS and MS degrees in computer science from Tianjin University, China. She is currently working toward the PhD degree in computer science at Worcester Polytechnic Institute. Her primary interests lie in the areas of intelligent tutoring systems, particularly, student modeling and educational data mining. She has also worked in the area of cognitive modeling and psychometrics. Her research has contributed to the design and evaluation of educational software, developed computing techniques to address problems in user learning, and produced basic results on the tracking student learning of mathematical skills. Her paper summarizes the current state of her thesis work on cognitive skill assessing in the ASSISTment project, on which she has been working since 2004.

Neil T. Heffernan received the summa cum laude degree in history and computer science from Amherst College, and the PhD degree in 2001 at Carnegie Mellon University. He is an associate professor of computer science at Worcester Polytechnic Institute and a creator of WPI's new Learning Sciences & Technology PhD program. For his dissertation, he built the first intelligent tutoring system that incorporated a model of tutorial dialog, and was one of the first intelligent tutoring systems on the Web. He does multidisciplinary research in intelligent tutoring systems, artificial intelligence, psychometrics, and cognitive science. He has more than 25 peer-reviewed publications and received about $9 million in funding on more than a dozen different grants.

Cristina Heffernan received the BS degree in mathematics from Lewis and Clark College, the MS degree in mathematics education from The University of Pittsburgh, and the MA degree in teaching from Towson State University. She serves as a cocreator of the ASSISTment system and the chief subject matter expert. Prior to that, she was a math coach and has seven years of experience teaching middle school math.

Murali Mani received the MS and PhD degrees from the University of California, Los Angeles, and the BTech degree from the Indian Institute of Technology, Madras. He is an assistant professor in the Department of Computer Science at Worcester Polytechnic Institute, where he joined in 2003. His research interests are in database systems, where he focuses on data integration, Web and XML systems, and data management for health informatics.