Feng, M, Heffernan, N., Heffernan, C. & Mani, M. (accepted). Using mixed-effects modeling to analyze different grain-sized skill models. Accepted by the IEEE Transactions on Learning Technologies Special Issue on Real-World Applications of Intelligent Tutoring Systems.

Author's note: This is not the final version. Please contact the first author before citing the paper.

Using Mixed-Effects Modeling to Analyze Different Grain-Sized Skill Models

Abstract

Most assessments, like the math subtest of the SAT or the GRE, are unidimensional, in that they treat all questions on the test as sampling a single underlying knowledge component (i.e., concept, procedure or skill). On the other hand, teachers want cognitively diagnostic information (Nichols, Chipman & Brennan, 1995) that they can use to inform their instruction. 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 information about fine- grained knowledge components? In this paper we present some encouraging results about our attempt to provide a fine-grained model for a United States state test. In step 1, a fine-grained skill model was developed by having content specialists review the state test items to identify their required skills. In step 2, we performed statistical analyses of the model based on data collected in two school-years' usage of an online tutoring system, the ASSISTment System. We show that our fine-grained model could improve prediction compared to other coarser-grained models, and an IRT-based unidimensional model. With that said we don't know a great deal about the validity of each individual knowledge construct; all we report is that in total, using the finer-grained model we can better predict state test scores, but we don't know which knowledge components are the ones that are doing a great job versus which ones are maybe not as valid as others.

Using Mixed-Effects Modeling to Analyze Different Grain-Sized Skill Models

Abstract

Most assessments, like the math subtest of the SAT or the GRE, are unidimensional, in that they treat all questions on the test as sampling a single underlying knowledge component (i.e., concept, procedure or skill). On the other hand, teachers want cognitively diagnostic information (Nichols, Chipman & Brennan, 1995) that they can use to inform their instruction. 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 information about fine-grained knowledge components? In this paper we present some encouraging results about our attempt to provide a fine-grained model for a United States state test. In step 1, a fine-grained skill model was developed by having content specialists review the state test items to identify their required skills. In step 2, we performed statistical analyses of the model based on data collected in two school-years' usage of an online tutoring system, the ASSISTment System. We show that our fine-grained model could improve prediction compared to other coarser-grained models, and an IRT-based unidimensional model. With that said we don't know a great deal about the validity of each individual knowledge construct; all we report is that in total, using the finer-grained model we can better predict state test scores, but we don't know which knowledge components are the ones that are doing a great job versus which ones are maybe not as valid as others.

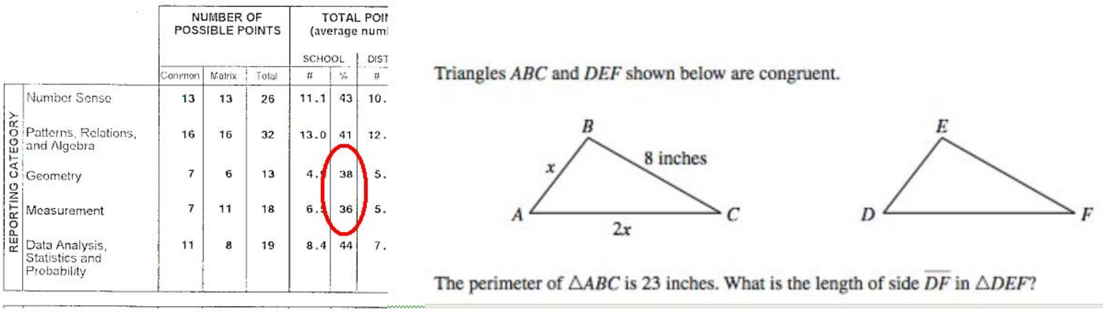
Keywords: ASSISTments, cognitive diagnostic assessment, fine-grained skill model, mixed-effects model

1. Introduction

1.1 Motivation

There is an increased interest in building cognitive diagnostic models (e.g. Griel, Wang & Zhou, 2008). Most large standardized tests (like the math sub-test of the Graduate Record Examination (GRE)) are what psychometricians call "unidimensional" in that they are analyzed as if all the questions are tapping a single underlying skill. It is this assumption of unidimensionality that makes computer adaptive testing possible for the GRE. However, cognitive scientists such as Anderson & Lebiere (1998), believe that students are learning individual skills, and might learn one skill but not another. 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 8th grade math Massachusetts Comprehensive Assessment System (MCAS[1]) test). Such uni-dimensional tests work pretty well at telling you which students are performing well but are not good at *informing educators* about how to help students.

[1] The Massachusetts Comprehensive Assessment System (MCAS) test is a state-administered standardized test that tests students in English, math, science and social studies for grades 3 to 10. We focused on only 8th grade mathematics. See http:// http://www.doe.mass.edu/mcas/ for more information about MCAS.



		NUMBER OF POSSIBLE POINTS			TOTAL POI (average num		
					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.

Triangles ABC and DEF shown below are congruent.

B
x
8 inches
A
2x
C
E
D
F

The perimeter of $\triangle ABC$ is 23 inches. What is the length of side $\overline{DF}$ in $\triangle DEF$?

(a) A school level report.

(b) Item 19 from the 2003 MCAS

Figure 1. A report showing low percent correct for Geometry and Measure and b) an MCAS item

The question of tagging items to learning standards is very important because schools (e.g. Worcester Public Schools in Worcester, Massachusetts) seek to use the MCAS assessments in a data-driven manner to provide regular and ongoing feedback to teachers and students on progress towards instructional objectives. For instance, the School Improvement Teams at each school review the results from the previous year to analyze which items their students performed particularly poorly on. However, teachers and parents also want better feedback than they currently receive. 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 5 mathematical reporting categories, known as “Strands.” And the state’s “Curriculum Framework” breaks the 5 strands into 39 individual “learning standards” for 8th grade math and tags each item with one of the 39 standards. The MCAS reporting system is representative of other states’ reporting systems.

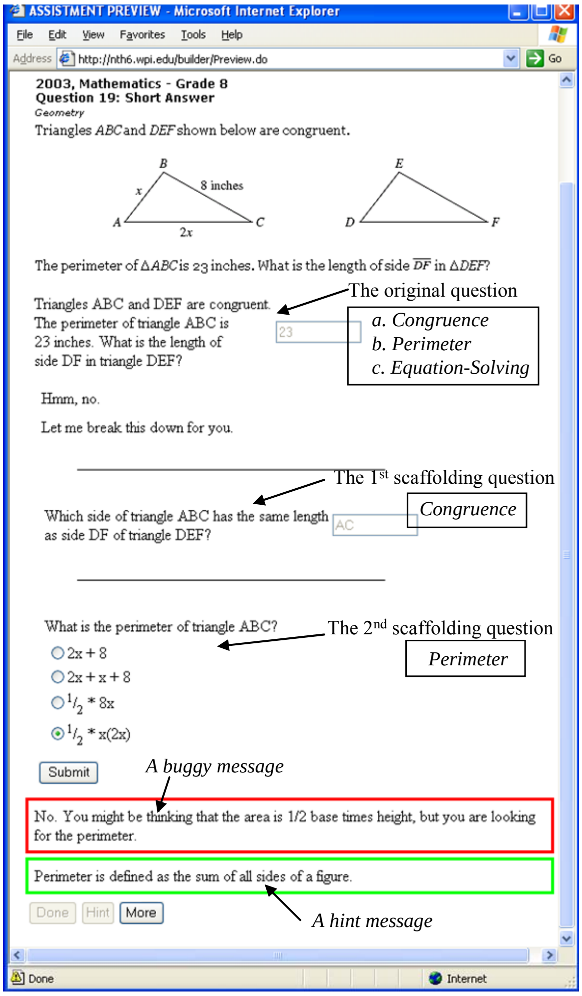
In 2004, a principal handed us a report, which he received from the state, and asked that we focus efforts on Geometry and Measurement because his students scored poorly in those areas (receiving 38% and 36% correct compared to over 41+% correct in the three other reporting categories, see Figure 1(a)). 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, Valenzuela, & Ortiz, 2002). Therefore, though, receivers of such reports are being told to be “data-driven” and use the reports to inform their instruction, the MCAS reports themselves are never designed to give feedback at a grain size that could be used for the purpose. A reader can get some intuition on why this is the case by trying item 19 from the 2003 MCAS shown in Figure 1(b). Then ask yourself “What makes this item difficult?” Clearly, this item includes elements from four of the 5 “strands” (only missing “Data Analysis, Statistics and Probability”). It is Algebra, Geometry (for its use of congruence), Number Sense (for doing the arithmetic operations), or Measurement (for the use of perimeter). Ignoring this obvious overlap, the state chose just one of the 5 strands to classify the item. It turns out, the state classifies it as Geometry and among the 39 learning standards “G.2.8-understanding-and-applying-congruence-and-similarity”, but below we will show how our methodology is creating evidence to suggest, as you might expect, that there is more to this problem than just 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. As a teacher has said “It does affect reports... because then the state sends reports that say that your kids got this problem wrong so they’re bad in geometry – and you have no idea, well you don't know what it really is – whether it's algebra, measurement, or geometry." It would be easier for a teacher to make data-driven changes in her classroom if she had a more detailed analysis of her students' learning. Students' weaknesses need to be addressed by helping them with concepts and skills that are neither too easy nor too hard.

We are engaged in an effort to investigate if we can do a better job of predicting a large scale test by modeling individual skills in a finer grain size. Griel, Wang & Zhou (2008) presented a study including building a cognitive model. They proposed two directions for future research, and one of the directions is to increase understanding of how to specify an appropriate grain size or level of analysis with a cognitive diagnostic assessment. In this paper, we consider four skill models with different granularity, including a unidimensional model and a fine-grained model developed at WPI (Section 2 describes how the model was built) with 78 skills. The four models are structured with an increasing degree of specificity as the number of skills goes up. The measure of model performance is the accuracy of the predicted MCAS test score based on the assessed skills of the students. What we refer to as a "skill model" is referred to as "Q-Matrix" by some Artificial Intelligence researchers (Barnes, 2005) and psychometricians (Tatsuoka, 1990); Croteau, Heffernan & Koedinger (2004) called it "transfer model"; while Cen, Koedinger & Junker (2005), and Griel, Wang & Zhou (2008) used the term "cognitive model". Researchers in educational measurement field such as Leighton & Gierl (2007, p.6) consider a cognitive model as a "simplified description of human problem solving on standardized educational task which helps to characterize the knowledge and skills students at different levels of learning have acquired and to facilitate the explanation and prediction of students' performance". 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. Given that the fine-grained model is composed of 78 skills, people might think the model would naturally fit the data better than the skill models that contain far less skills. Moreover, they may even worry that we were overfitting our data by fitting a model with so many free parameters. However, we were not evaluating the effectiveness of the skill models over the same online student data (collected by the ASSISTment system) based on which models will be constructed. Instead, we used totally different data (from MCAS test, the external, paper-and-pencil based state test) as the testing set. We evaluate our models using the 8th grade 2005 test, which we will refer to as the state test. 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.

1.2 The ASSISTment Project

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). For instance, the high-stakes MCAS test is a graduation requirement in which all students educated with public funds are required to participate. It administers rigorous standardized tests in English, math, history and science in grades 3–10 every year. Students need to pass the math and English portions of the 10th grade versions in order to get a high school diploma without further remediation. In 2003, a full 10% of high school seniors were predicted to be denied a high school diploma due to having failed to pass the test on their fourth try. Moreover, the state of Massachusetts has singled out student performance on the 8th grade math test as an area of highest need for improvement[2].

There is a large interest in "Formative Assessment" in K-12 Education (Olson, 2004) with many companies[3] providing such services. Some teachers make extensive use of practice tests and released test items to help identify learning deficits for individual students and the class as a whole. However, such formative assessments not only require great effort and dedication, but they also take valuable time away from instruction. Some online testing systems (such as Renaissance Learning[4]) automatically grade students and provide reports but they may not be informative as they do not maintain sufficiently rich data records for students and therefore cannot report on a fine-grained model of student knowledge.


ASSISTMENT PREVIEW - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://nth6.wpi.edu/builder/Preview.do
Go
B
E
x
8 inches
A
C
2x
D
F
Submit
Done
Hint
More
Done
Internet

2003, Mathematics - Grade 8
Question 19: Short Answer
Geometry

Triangles ABC and DEF shown below are congruent.

The perimeter of ΔABC is 23 inches. What is the length of side $\overline{DF}$ in ΔDEF?

The original question

a. Congruence
b. Perimeter
c. Equation-Solving

Triangles ABC and DEF are congruent. The perimeter of triangle ABC is 23 inches. What is the length of side DF in triangle DEF? 23

Hmm, no.

Let me break this down for you.

The 1st scaffolding question

Congruence

Which side of triangle ABC has the same length as side DF of triangle DEF? AC

The 2nd scaffolding question

Perimeter

What is the perimeter of triangle ABC?

○ 2x + 8
○ 2x + x + 8
○ $^1/_2$ * 8x
● $^1/_2$ * x(2x)

A buggy message

No. You might be thinking that the area is 1/2 base times height, but you are looking for the perimeter.

Perimeter is defined as the sum of all sides of a figure.

A hint message

Figure 2. An ASSISTment shown student working on an ASSISTment

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. Yet, traditionally these two areas of testing (i.e. Psychometrics) and instruction (i.e., math educational research and instructional technology research) have been separate fields of research with their own goals. A solution must involve a way whereby students can take an assessment, but also learn as they are being assessed; unfortunately statisticians have not done a great deal of work to enable assessment of students while they are learning during the test.[5] A solution needs to be found so that teachers can get the benefit of being data-driven in trying to meet instructional objectives, but at the same time, make sure that their students' time is spent primarily on learning. To help resolve this dilemma, the U.S. Dept of Education funded Heffernan and Koedinger to build a web-based tutoring system that would also do assessment at the same time. Assistance and assessment are integrated in the system ("ASSISTment"[6]) that will offer instruction to students while providing a more detailed

[2] http://www.doe.mass.edu/mcas/2002/results/summary.pdf

[3] Including nwea.org/assessments/, measuredprogress.org, Pearson and ww.cddre.org/Services/4Sight.cfm

[4] www.renlearn.com

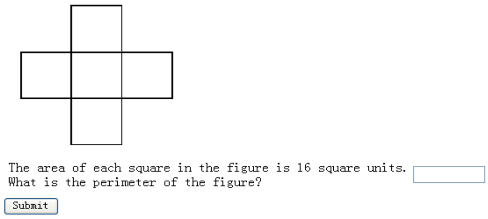
[5] Standard psychometric models (e.g. van der Linden & Hambleton, 1997) assumes the amount of learning happens during a test is limited. Some works have been done to measure growth and change (e.g. Tan, Imbos & Dos, 1994; Embretson, 1992; Fischer & Seliger, 1997), but they are not based on testing data where students are actively learning materials.

[6] The term "ASSISTment" was coined by Kenneth Koedinger and blends **assist**ing and assess**ment**.

evaluation of their abilities to the teacher than is possible under current approaches. Unlike other assessment systems, the ASSISTment technology also provides students with intelligent tutoring assistance while the assessment information is being collected.

The ASSISTment system is a tutoring program and each week when students work on the website, the system "learns" more about the students' abilities and thus, it can hypothetically provide increasingly accurate predictions of how they will do on a standardized mathematics test. It helps students to work through tough problem by breaking the problem into steps; meanwhile, it collects data related to different aspects of student performance such as accuracy, speed, help-seeking behavior and attempts as students interact with the system. Recent studies showed that ASSISTments can assess students accurately (Feng, Heffernan & Koedinger, 2006a, 2006b; Feng, Beck, Heffernan, & Koedinger, 2008) as it assists student learning (Razzaq & Heffernan, 2006, 2007; Feng, Heffernan, Beck & Koedinger, 2008). Besides, Mendicino, Razzaq & Heffernan (in press) showed that students learned significantly more when given home in ASSISTment system than when doing traditional paper-and-pencil homework. Based on the rich source of data, we were also able to report continuously via various reports (e.g. Feng & Heffernan, 2007) to teachers and other stakeholders to help them better understand students' performance and progress.

In Massachusetts, the state department of education has released 10 years (1998-2007) worth of 8th grade MCAS test items, almost 400 items, which we have turned into ASSISTments by adding "tutoring". Each ASSISTment 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 by our content experts to coach students who fail to answer the original question. Item 19 of the 2003 MCAS is shown in Figure 1. An ASSISTment that was built for this item is shown in Figure 2 which shows the state of the interface when the student is partly done with the problem. The first scaffolding question appears only if the student gets the item wrong. We see that the student typed "23" (which happened to be the most common wrong answer for this item from the data collected). After an error, students are not allowed to try the item further, but instead must then answer a sequence of scaffolding questions (or "scaffolds") presented one at a time. Students work through the scaffolding questions, possibly with hints, until they eventually get the problem correct. If the student presses the hint button while on the first scaffold, the first hint is displayed, which would be the definition of congruence in this example. If the student hits the hint button again, the second hint appears which describes how to apply congruence to this problem. If the student asks for another hint, the answer is given. Once the student gets the first scaffolding question correct (by typing "*AC*"), the second scaffolding question appears. Buggy messages will show up if the student types in a wrong answer as expected by the author. Figure 2 shows a buggy messages that appeared after the student clicked on "½*x(2x)" suggesting he might be thinking about area. Once the student gets this question correct he will be asked to solve 2x+x+8=23 for 5, which is a scaffolding question that is focused on equation-solving. So if a student got the original question wrong, what skills should be blamed? This example is meant to show 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 will be able to tell if the student got the question wrong because they did not know congruence versus not knowing perimeter, versus not being able to set up and solve the equation. As a matter of logging, the student is only marked as getting the item correct if they answered the questions correctly before asking for any hints or encountering scaffolding.



The area of each square in the figure is 16 square units.
What is the perimeter of the figure?

Figure 3. The original question of item 27 of 1999 MCAS test

Figure 3 shows the original question of another ASSISTment built for item 27 of 1999 MCAS test. The ASSISTment provides two scaffolding questions. The first one asked "What is the length of one side of a square in the figure?" and the second says "Now you have enough information to find the perimeter of the figure. What do you think it is?" In the "WPI-78", the original question was tagged with 2 skills: "Perimeter" and "Area"; the first scaffolding question is associated with "Perimeter" and the second one "Area".

In the first year the ASSISTment system was launched, the 2004-2005 school year, some 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. Since then the number of users has expanded every year and more than 3000 students from Massachusetts used the system during the school year of 2007-2008.

1.3 Literature Review

Modeling student response data from intelligent tutoring systems has a long history (Corbett, Anderson, & O'Brien, 1995; Draney, Pirolli, & Wilson, 1995). Corbett and Anderson did show that they could get better fitting models to predict student performance in LISP programming by tracking individual production but their system never asked questions that were tagged with more than one production, which is the sort of data we have (described below). Our collaborators (Ayers & Junker, 2006) are engaged in trying to allow multi-mapping using a version of the WPI-78 but report their Linear Logistic Test Model (LLTM) does not fit well. A "multi-mapping" skill model, in contrast to a "single-mapping" or a "non-multi-mapping" model, allows one item to be tagged with more than one skill. Anozie & Junker (2006), are looking at this same data set, also trying to predict the same state test scores we will describe below, but they are not using skills at all, and since their method is unidimensional, in one sense representing the more traditional psychometric approach. This paper will not be able to compare the results of these different approaches and models, as we are all using slightly different versions of the same data set.

Others, in the psychometrics field, have developed multi-dimensional Item Response Theory models but these models have generally not allowed multi-mapping. These models permit student performance to be measured by comparisons within items. For instance, Bock, Gibbons, and Muraki (1988) developed a multidimensional IRT model that identifies the dimensions that are needed to fit test data, similar to an exploratory factor analysis. Though different approaches have been adopted to develop skill models and thus model students' responses, as far as we know, little effort has been put in that compares different grain-sized skill models in the intelligent tutoring system area. The few that have done this have done so in a non-multi-mapped manner (Corbett, Anderson, & O'Brien, 1995; Draney, Pirolli, & Wilson, 1995). While we come to this work from the point of view of an intelligent tutoring system's researchers, in the education field more broadly, researchers want to fit students' data collected in the traditional paper-and-pencil method. Unfortunately, the only work we are aware of that shows that by building fine-grained skill models researchers could build better fitting models is by Yun, Willett and Murnane (2004). Yun et al. (2004) developed an alternative curriculum framework by examining questions in the MCAS state test and they performed confirmatory factor analysis on students' item level response data from the 2001 MCAS English language arts (ELA) test to investigate how well the MCAS test items map onto the state learning standards' and onto the alternative framework. Their result showed that the alternative framework fits student response data better as measured by Akaike Information Criteria (AIC), suggesting the state's learning standards for ELA is subject to improvement.

Recently, Mislevy (2006) described six steps in model-based reasoning in science. These steps, including model formation, model elaboration, model use, model evaluation, model revisions and model-based inquiry, provide a framework for considering our progress in developing & refining cognitive models. Following these steps, the rest of the article 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. And we conclude our work in Section 4 and bring up the issue of model refinement and model-based inquiry as part of our future work.

2 Towards a Fine Grained Skill Model

2.1 Developing a fine grained model for 8th grade MCAS

In April, 2005, we staged a seven hour long "coding session", where we invited our subject-matter expert, Cristina Heffernan, with the assistance of the second author to create a set of skills and to use those skills to tag all of the existing 8th grade MCAS items with these skills[7]. The process is quite similar to the study conducted by Gierl, Wang, & Zhou (2008). There were about 300 released test items for us to code. Because we wanted to be able to track learning between items, we wanted to come up with a number of skills that were somewhat fine-grained but not too fine-grained such that each item had a different skill. We therefore imposed upon our subject-matter expert that no one item would be tagged with more than 3 skills[8]. It is not coincidence that many of our assistments have about three scaffolding questions (The question in Figure 2 shows two scaffolding questions and there are two more after that, for a total of four scaffolding questions for that ASSISTment); we wanted the fine grained-ness of the modeling to match the fine grained-ness of the scaffolding. We knew we wanted most of our scaffolds to have identifiability, meaning that each scaffolding question should be tagged with only one skill. We wanted identifiability because we thought that when a student got a question wrong, that was tagged with two skills, we should have a very hard time coming up with a method that would be able to blame the "correct" skill. Identifiability of scaffolds avoids those problems, but of course forces the modeler to have to use fewer skills.

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 ASSISTments, there was 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

[7] We hand-coded the skills in this work. Though, others have done work showing it is possible to use an automatic technique such as LFA (Cen, Koedinger & Junker, 2005), Q-matrix (Barnes, 2005) method for topic construction, rule space method (Tatsuoka, 1990). We choose to use a hand-coding method, as it is easier to interpret the meaning of skills when coded by hand.

[8] This April 2005 tagging session was preceded by another tagging session where we did not impose any constraints; we wound up with some questions being tagged with 9 skills, and we thought it would have been impossible to build a good fitting model.

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 all questions tagged with the same set of skills have the same degree of difficulty.

Figure 4. Picture showing the 3rd author at far left, working with staff to develop the model. Each table has rows of items, organized by skill.

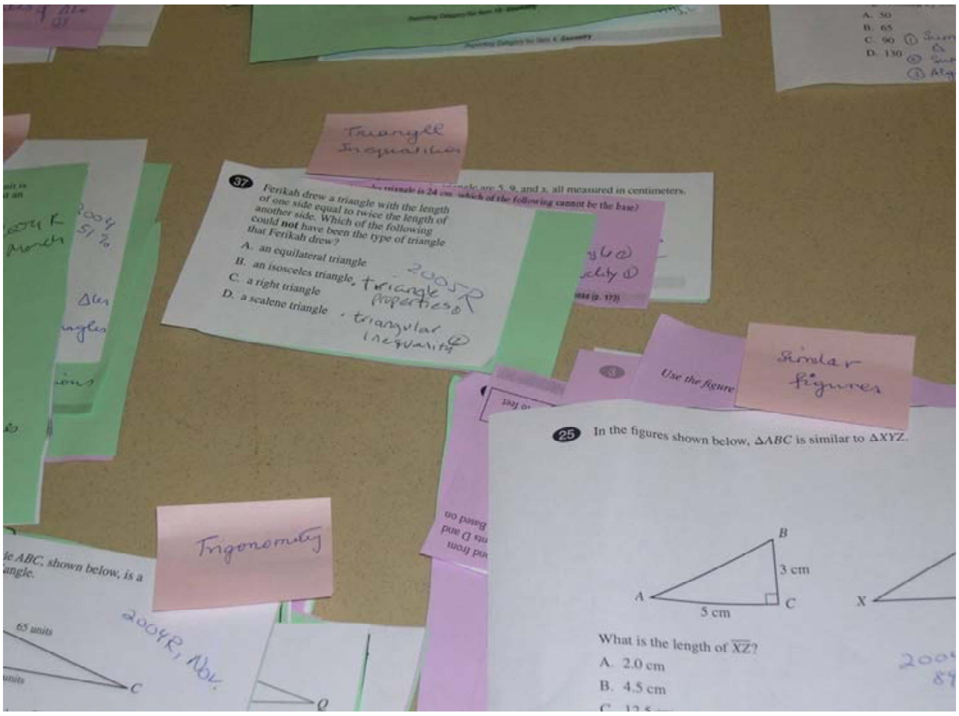

Triangle Inequalities
37
Ferikah drew a triangle with the length of one side equal to twice the length of another side. Which of the following could not have been the type of triangle that Ferikah drew?
A. an equilateral triangle
B. an isosceles triangle
C. a right triangle
D. a scalene triangle
2005R
triangle properties
triangular Inequality
Similar figures
Use the figure
25
In the figures shown below, ΔABC is similar to ΔXYZ.
A
B
C
X
3 cm
5 cm
What is the length of XZ?
A. 2.0 cm
B. 4.5 cm
Trigonometry
2004R, Nov.
65 units
C

Figure 5. The April 2005 coding session when all the existing MCAS items are tagged with skills

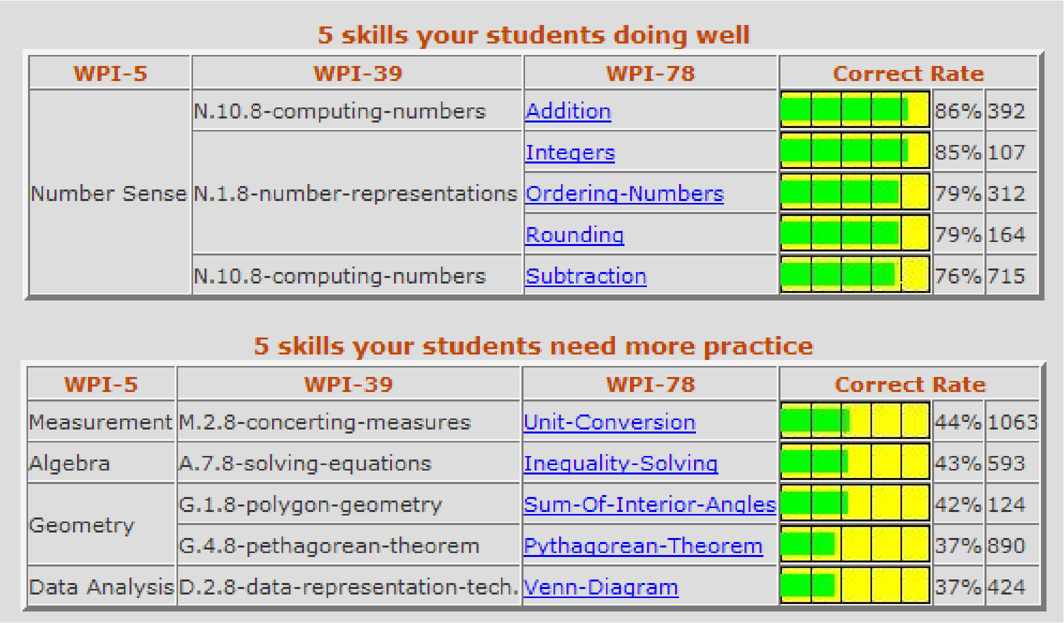
During the "coding session", the subject-matter expert was free to create whatever skills she thought appropriate. We printed three copies of each item so that each item could show up in different piles, where each pile of printed items represented a skill. We reviewed the items, solved the problems and conducted cognitive task analysis to identify what is the knowledge needed to perform each task. Though we have English names for the skills, those names are just a handy tag; the real meaning of a skill must be divined from the questions with which it is associated. The name of the skill served no-purpose in our computerized analysis. When the coding session was over, we had six tables covered with 106 piles of items (See Figure 4 & Figure 5). We wound up with about 106 skills, but not every skill that was created was eventually involved in the data source 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" and 5) "Measurement". The Massachusetts Department of Education actually tags each item with exactly one of the 5 strands, but our mapping was not the same as the state's mapping. Therefore, it was named WPI-5. Furthermore, we allowed multi-mapping, i.e. allow 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 5 strands) in the Massachusetts Curriculum Framework, associated each skill in WPI-78 to one of the learning standards, and, thus, created the model WPI-39. After the students had taken the 2005 state test, 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. After our 600 students had taken the 2005 state test and 2006 state test, the state released the items from that test, and we had our subject matter expert tag up the items in that test as well.

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 WPI-1, WPI-5 and WPI-39 models are 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 10 of the 78 skills in the WPI-78 skill model. In the second column we see how the 5 skills in WPI-78 are nested inside of “Patterns, Relations and Algebra”, which itself is one piece of the 5 skills that comprise the WPI-5 skill model.

Consider item 19 from the 2003 MCAS test (as shown in Figure 2). In the WPI-78 skill model, the first scaffolding question is tagged with “congruence”, the second tagged with “perimeter”, the third tagged with “equation-solving”. 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 therefore tagged correspondingly with “Geometry”, “Measurement” and “Patterns, Relations and Algebra”, and just one skill of “math” at the WPI-1. Similarly, the original question of item 27 from the 1999 MCAS test shown in Figure 3 is tagged “Perimeter” and “Area”, and therefore it is tagged with “Measurement” in the WPI-5, and then again “math” in the WPI-1.



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

Figure 6. A skill report showing skills on which students performed well or poorly

One might wonder what else developing these skill models are useful for. Most intelligent tutoring systems require the construction of complex models that represent student knowledge states used to track student knowledge. If we have a better skill model we should be able to do a better job of predicting which items students will get correct in real-time. That means we should be able to do a better job of selecting the next best item for them to work on. In our tutoring system, the next best item will 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 help to address the issues as we mentioned in section 1.1 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 should be able to 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 that teachers can get using the ASSISTment system is shown in Figure 6. Teachers can see how their students are doing on each skill and can determine where they need to spend the most time in their classroom instruction.

2.2 Validating the fine-grained model

To validate the fine-grained model, a study was conducted to see how different educators agree on the tagging of problems and to calculate an inter-rater reliability. We randomly selected 50 sample items from the ASSISTment problem pool, and asked our subject-manner expert to give a short description of the skills in case the names of skills are not self-explanatory. Two middle school math teachers (we will refer to them as teacher A and teacher B) were given the items and the list of skills, and were asked to tag the items with proper skills individually. The skills were organized in the hierarchical structure as shown in Table 1 to help teachers locate skills. It took each teacher around 30 to 45 minutes to finish the whole tagging process on paper, which is faster than what we would have expected. Yet, this is not surprising with regard to the fact that both teachers are experienced with the ASSISTment system and familiar with the skills in the fine-grained model since they have been building new items and tagging their items (not included in the sample) with the skills using the ASSISTment builder (as described in the next section).

We first looked at the results to see how much homogeneity there is in the tagging made by the two teachers and our subject-matter expert at item level. Since we allow multi-tagging and imposed no restriction on how many skills can be tagged to an item, there are many ways that the tagging can differ among different raters. In one case, teacher A tagged an item with 3 skills, teacher B tagged the item with one skill (one of the three skills used by teacher A), and in our model, the same item was tagged with 2 skills (a subset of the skills used by teacher A); in another case, teacher A picked the same skill as in our model, yet teacher B picked another skill; etc. For instance, the item shown in Figure 7 is tagged with “Supplementary-angles” and “Traversals” in our model, yet both of the two teachers tagged it with only the skill “Traversals”.

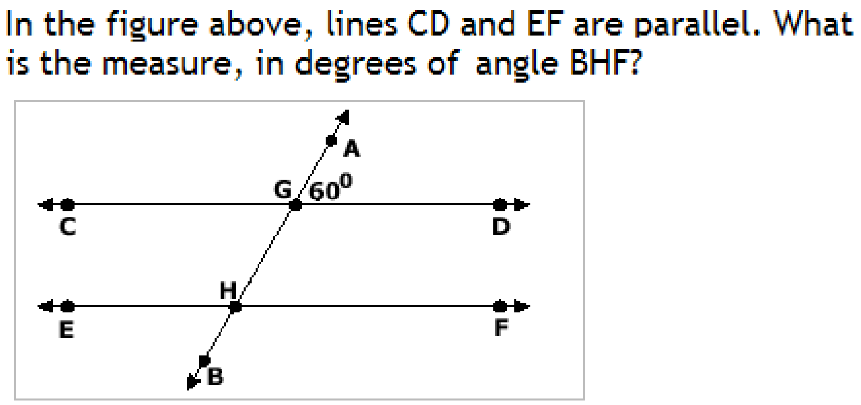

In the figure above, lines CD and EF are parallel. What is the measure, in degrees of angle BHF?

Figure 7. A sample problem where the two teachers partially agreed with our model on the skill tagging.

We considered there is a full agreement between two teachers and our subject-manner expert at the item level when both of two teachers tagged an item with exactly the same set of skills as being used in our skill model. A partial agreement on an item means at least one rater agreed with another on at least one skill tagging. A disagreement means neither of the two teachers tagged an item with any skill that was associated with that item in our model. Among the 50 items in the sample, the three raters fully agreed on 11 items (22%), partially agreed on 28 items (56%), and disagreed on 11 items.

Inter-rater reliability was also assessed by comparing the tagging between any pair of raters (teachers A/B or our subject-matter expert) at skill level. We examined all the skills that were selected to tag to an item by two raters one by one to see how often two raters agreed/disagreed with each other. The results suggested raters disagreed with each other more often than they agreed at skill level, although the difference is not always reliable as indicated by a modified sign-test[9]. For instance, overall teacher A and our subject manner expert agreed on 38 skills but disagreed on 54 skills while teacher B agreed with our subject manner expert on 37 skills but disagreed on 71[10] skills.

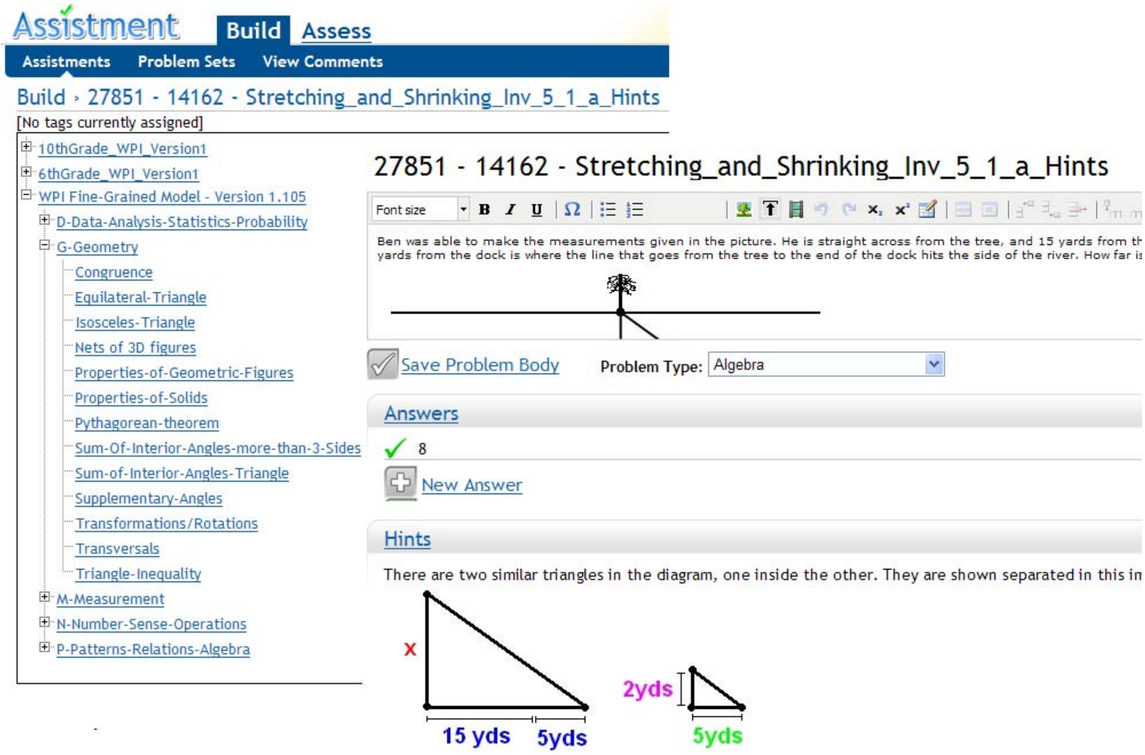
All in all, we did not find a lot of concordance between the teachers and our subject manner expert's tagging. This disappointed us a bit, honestly, but it also led us to believe more that model development and skill tagging are difficult tasks. With 106 skills in the model, it is hard for people to pick up one, agree on one skill (for two raters, the chance is at one out of 10,000, although for most skills, experience teachers can quickly tell if an item is or is not associated with the skill), even harder when we allow multi-mapping. In Section 3, we will evaluate our model on how well our skill model helps track student performance, even though the teachers did not seem agree with us very much.

2.3 Tagging skills to problems in ASSISTments

The ASSISTment Builder (Heffernan et al., 2006) has the authoring tools that mainly support content creation in the ASSISTment system. In addition to content creation, it allows content authors to tag the ASSISTments they have built with skills, which are organized into sets known as skill models. This tool, shown in Figure 8, provides a means to map certain skills to specific problems to specify that a problem involves knowledge of that skill. This mapping between skills and problems allows the reporting system (Feng & Heffernan, 2007) to track student knowledge over time using longitudinal data analysis techniques. In Feng, Heffernan & Koedinger (2006a), we report on the ability to track the learning of individual skills using the coarse-grained model, WPI-5 that classifies each 8th grade MCAS math item in one of five categories: Algebra, Measurement, Geometry, Number Sense, and Data Analysis. The skills are organized in a hierarchical structure as shown in Figure 5. The user is allowed to browse the skills within each transfer model and to map the ones they select to a problem.

[9] http://www.fon.hum.uva.nl/Service/Statistics/Sign_Test.html

[10] In general, teacher B tends to tag an item with more skills than teacher A and our subject manner expert. This is why when we compared tagging by teacher B and our subject manner expert, there are more skills for us to look at than when we compared teacher A's tagging to our subject manner expert's tagging.



Build › 27851 - 14162 - Stretching_and_Shrinking_Inv_5_1_a_Hints

[No tags currently assigned]

- 10thGrade_WPI_Version1
- 6thGrade_WPI_Version1
- WPI Fine-Grained Model - Version 1.105
  - D-Data-Analysis-Statistics-Probability
  - G-Geometry
    - Congruence
    - Equilateral-Triangle
    - Isosceles-Triangle
    - Nets of 3D figures
    - Properties-of-Geometric-Figures
    - Properties-of-Solids
    - Pythagorean-theorem
    - Sum-Of-Interior-Angles-more-than-3-Sides
    - Sum-of-Interior-Angles-Triangle
    - Supplementary-Angles
    - Transformations/Rotations
    - Transversals
    - Triangle-Inequality
  - M-Measurement
  - N-Number-Sense-Operations
  - P-Patterns-Relations-Algebra

27851 - 14162 - Stretching_and_Shrinking_Inv_5_1_a_Hints

Ben was able to make the measurements given in the picture. He is straight across from the tree, and 15 yards from th
yards from the dock is where the line that goes from the tree to the end of the dock hits the side of the river. How far is

Problem Type: Algebra

Answers

✓ 8

Hints

There are two similar triangles in the diagram, one inside the other. They are shown separated in this in

Figure 8. A screen shot showing an item and a list of skills that could be tagged to the item in the ASSISTment builder

3 Statistical Analysis of the Skill Model

3.1 Data Source

We collected data from 447[11] students who used our system from Sep. 17, 2004 through May 16, 2005 for on average 7.3 days (one period per day)[12]. All these students had worked on the system for at least 6 days (one session per day). We excluded data from the students' first day of using the system considering the fact that 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), includes both responses to original questions and to scaffolding questions. It contains about 138 thousand data points, among which around 43 thousand 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 school year as well. The data set involves 474 students who on average worked in the system for

[11] The amount of data is limited by the maximum memory allowed by the open source statistical package we used.

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

Table 2. Sample Raw Data

RowID	StudentID	State Test ID	ItemID	WPI-78 skills	Original?	Response	Month Elapsed
1	950	2003-#19	326	Congruence	Y	0	1.32
2	950	2003-#19	326	Perimeter	Y	0	1.32
3	950	2003-#19	326	Equation-Solving	Y	0	1.32
4	950	2003-#19	327	Congruence	N	0	1.32
5	950	2003-#19	328	Perimeter	N	1	1.32
6	950	2003-#19	329	Equation-Solving	N	0	1.32
7	950	2003-#19	330	Equation-Solving	N	0	1.32
9	950	1999-#27	1183	Perimeter	Y	0	2.94
10	950	1999-#27	1183	Area	Y	0	2.94
11	950	1999-#27	1184	Perimeter	N	1	2.94
12	950	1999-#27	1185	Area	N	1	2.94

5.5 days and answered 51 original questions. The item level response data from the 2006 MCAS tests is available for these students too. This data set will be referred to as Data-2006.

Both of the data sets are organized in the way that 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 2 skills in a model, then for each response made to the question there would be 2 rows in the data set, with skill names listed in a separate column. Students' 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 answer was given. Additionally, a column is associated with each response, indicating the number of months elapsed since September 17, 2004 till the time when the response was made. Thus the number of months elapsed for a response made on September 17th will be zero, and the number will be 1 for a response made at October 17th, 2004, and so on. This gives us a longitudinal, binary response data set across the school year.

Table 2 displays 12 rows of the raw data for one student (system ID = 950) who finished the item 19 (shown in Figure 1) and item 27 (shown in Figure 2) on two different days. The first 7 rows represent the student' response on item 19 (with original item ID[13] being 326) and the remaining 6 rows show his response on item 27 (with original item ID being 1183). We can see that since the original question of item 19 was tagged with 3 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 2 skills as shown in row 9 and row 10. For both items, the student answered the original questions incorrectly (indicated by "0" in the response column of rows 1-3 and rows 9-10) and thus 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), but failed to answer all of the other scaffolding questions. In contrast, although the student did not get item 27 right on the first shot, the student went through both scaffolding questions correctly. WPI-78 is the skill model being used here.

[13] The "itemID" is a number that we used internally in the system to uniquely identify a question. It is displayed only for the purpose of interpreting the data.

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

We fit a longitudinal model, the mixed-effects logistic regression model, to our data to obtain an estimate of student knowledge on individual skills at a certain time, assuming student knowledge was changing linearly over time. 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. Snijders & Bosker (1999, Chapter 14) provide a practical summary of the mixed-effects logistic regression model and various procedures for estimating its parameters. Hedeker & Gibbons (in progress, Chapter 9) describes mixed-effects models for binary data that accommodate multiple random effects. As these sources indicate, the mixed-effects logistic regression model is a very popular and widely accepted choice for analysis of dichotomous data. It describes the relationship between a binary or dichotomous outcome and a set of explanatory variables. In this work, we adopted this model and fitted it to our longitudinal, binary response 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 for the *i*th 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:

$$p_i = \Pr(R_i = 1) = \frac{\exp(x_i^{'}\beta)}{1 + \exp(x_i^{'}\beta)}$$

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

$$\log[\frac{p_i}{1 - p_i}] = x_i^{'}\beta$$

In logistic regressions, the *logit* is called the link function because it maps the (0, 1) range of probabilities unto (-∞, +∞) 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.

Now we generalize the simple logistic regression model to 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 2-level representation of the model in terms of *logit* can be written as

Level-1 (or within-person) model:

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

Level-2 (or between-person) model:

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

Where

p_{ij} is the probability of a positive response for student *i* at time *j*

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" (Singer & Willett, 2003) since TIME is introduced as a predictor of the response variable, which allows us to investigate change over time. The fact that this mixed-effect regression model is linear in terms of *logit* enables us to fit generalized linear mixed models on the data in R (R Development Core Team, 2007), an open-source statistical environment. The models were fitted in R using *lmer()* function and "logit" was used as the link function. Below in the box is the sample code we ran in R to train a mixed-effects logistic regression model using both TIME and WPI-5 skills as covariates.

```
## train the model, using WPI-5 skill model
>> model.growth.WPI.5 <- lmer(response ~
monthElapsed + skills + skills * monthElapsed +
(monthElapsed | studentID), data= WPI.5, family=
binomial (link="logit"), control = list(msVerbose = 1,
usePQL = FALSE))
## extract the fixed effects of the model
>> fix.ef.WPI.5 <- fixef(model.growth.WPI.5)
## extract random effects for every student
>> ran.ef.WPI.5 <- ranef(model.growth.WPI.5)
```

After the model was constructed, the fixed-effects for the whole group (i.e. β_0, β_1 in the above 2-level model) and the random effects for each student (i.e. v_{0i}, v_{1i}) were extracted and then the two learning parameters "intercept" and "slope" (i.e. b_{0i} and b_{1i} in the model above) was calculated for each individual student (and for each skill if skill was introduced as factor into the model). Given this, we thus can apply the model on the items in the state test to estimate students' response to each of them.

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 4 skill models by our subject matter expert[14]. 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 come up with a prediction of 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

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

us the lowest probability among all the skills that apply to the item[15] 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[16].

The prediction functions we build 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 is suggested by Brian Junker, a statistician from Carnegie Mellon University (Junker, 2007).

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

where *MCAS*$_i$ is the actual MCAS score of the i$_{th}$ student, and *prediction*$_i$ is the predicted score from the prediction function being evaluated. For every model, we subtract 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 calculate 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 5 open response, 4 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 (We will refer to as RQ1): Would adding response data to scaffolding questions help us do a better job of tracking students' knowledge and thus more accurately predicting state test scores, compared to only using the original questions? Because the scaffolding questions break the test questions down into a series of simpler tasks that directly assess fewer knowledge components, we believe the ASSISTment system can do a more

[15] We admit that there are other approaches dealing with multi-mapped items. For instance, one way can be taking into consideration the conjunctive relationship among the skills and "somehow" combining the probabilities together to produce a "final" probability for the item. Using Bayesian Networks is also a reasonable way to deal with this situation and our colleagues Pardos, Heffernan, Anderson and Heffernan (2006) use this approach and seem to be getting similar results that fine-grained models enable better predictive models.

[16] We think 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), that 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.

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 4.

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 which skills are the real obstacles that prevent students from correctly answer the original questions. And 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/she 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 multi-mapping 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 do a better job of tracking students' knowledge and thus 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 3 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% after scaffolding questions were introduced; for WPI-5, the decrease is 1.21%; and the decrease of "% Error" is 2.88% for the WPI-39 and 5.79% 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 which 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 do a better job of dealing with credit-and-blame problems. We admit that here we have confounded the impacts of simply adding in scaffolding questions response data and adopting the credit-and-blame strategies. And we want to investigate their effects separately in the near future. 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.

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

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 items were tagged by the same two skills, which means we don't 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. Thus, the %Error of 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% on the WPI-78 seem impressive or poor? What is a reasonable goal to shoot for? Zero percent error? In Feng, Heffernan & Koedinger (2006b) we reported on a simulation of giving two MCAS tests in a row to the same students and then used one test to predict the other and got an approximate 11% error rate, suggesting that a 12% error rate is looking somewhat impressive.

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

Research Question 2 (We will refer to as 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 4 different skill models. As shown in Table 4, the WPI-78 had the best result, followed by the WPI-39, WPI-5, and followed by the 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.

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=.006
WPI-5	4.37	[4.07, 4.66]	12.85%	p<.0001
WPI-39	4.22	[3.94, 4.50]	12.41%	p=.21
WPI-78	4.11	[3.84, 4.39]	12.09%	p=.10
IRT-2005	4.36	[4.04, 4.68]	12.82%	
		DATA-2006		
WPI-1	6.58	[6.18, 6.99]	19.37%	p<.0001
WPI-5	6.51	[6.11, 6.90]	19.14%	p<.0001
WPI-39	4.83	[4.56, 5.11]	15.10%	p=.0001
WPI-78	4.99	[4.71, 5.28]	14.70%	p=.03
IRT-2006	4.67	[4.34, 4.99]	13.7%	

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 = .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<.001 in all cases), and WPI-1 is statistically reliably worse on predicting MCAS scores than the other models (p <.0001). This suggested that using finer-grained skill models was helpful in tracking students’ knowledge over time.

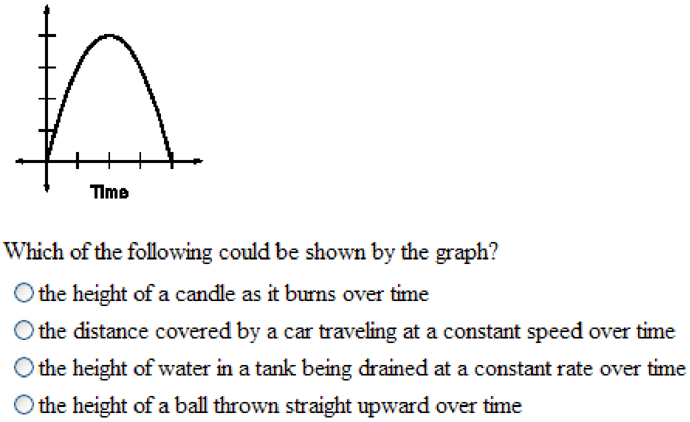
We want to stress that the main goal of this paper is to see if finer-grained skill models track students’ knowledge better and we claim the goal was achieved based on the result presented in Tables 3 & 4. Therefore, though questions such as “Are the improvements in accuracy from 4.42 to 4.11 meaningful?”, “What is the practical value of this improvement?” are interesting, they are beyond the scope of this paper. Besides, our results on student performance prediction are by no means the best. As a matter of fact, we trained an Item Response Theory (IRT; e.g. van der Linden & Hamilton, 1997) model that has been widely used in traditional testing area by psychometricians as a control. We fit the simplest model, the Rasch model (Fischer and Molenaar, 1995) that models student *i*'s dichotomous response (0 = wrong, 1 = 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 every item in MCAS has an average difficulty (β=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%, marginally higher than that of the WPI-78, 12.09% (p = .10). Yet, the Rasch model did in the next year where the %Error (13.70%) is reliably higher (p = .03) than that of the WPI-78 (14.70%).

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 in the internal fit. We do not bother to report the internal fit of our models using measures like BIC, because "we don't know how to". Less flippantly, the size of the datasets is different using the different models; the finer-grained models add additional rows for all questions that are tagged with more than one skill and BIC only make sense when the data is meant to be the exact same.

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 don't 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, and thus there is a tradeoff between the number of skills you would like, and the precision in the skill estimates. Possibly, one reason most tests like the GRE are unidimensional is that they might not ask enough questions to justify the additional fit they might get.

Comparing the results 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 vs. 7.3 sessions and 87 problems done). Additionally, the problem set administered to students in the two years are not the same either.

3.5.2 How well does the Model WPI-78 Fit the Data?

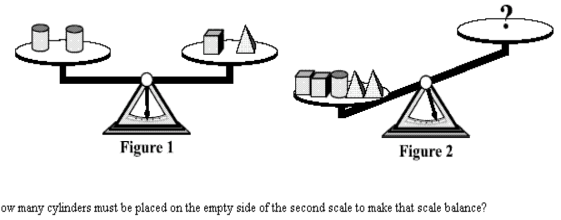


Which of the following could be shown by the graph?

○ the height of a candle as it burns over time
○ the distance covered by a car traveling at a constant speed over time
○ the height of water in a tank being drained at a constant rate over time
○ the height of a ball thrown straight upward over time

Figure 9. A question tagged with the skill "Qualitative-Graph-Interpretation"

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 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 Figure 9). 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 "The sum of the interior angles in a triangle is 180" clearly indicating that this was a skill that teachers were focused on teaching. Attentively, the skill that showed the least learning was called "Equation-Concept" (as shown in Figure 10). Out of the 78 skills, 7 coefficients predicted un-learning (i.e. the slopes are negative, which is presumably an issue of overfitting). In the future, we will investigate automating the process to remove such skills from the model and to re-fit the data.



How many cylinders must be placed on the empty side of the second scale to make that scale balance?

Figure 10. A question tagged with the skill "Equation-Concept"

Speaking of the accuracy of fit, we noticed 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" (correct rate around .5 ~.6)

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 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; 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"). We speculate this refining of this model using computational techniques like Learning Factors Analysis (Cen, Koedinger & Junker, 2005) could substantially improve the fit of this model to the data. This refinement *might* work better if students' learning of equation-concept should transfer to increase their performance on equation-solving and vice-versa.

All in all, we make no claim that the Q-matrix (Barnes, 2005) we created represented the best fitting Q-matrix possible. Nevertheless, we stand by the claim that this model, taken in total, is good enough that it can produce better fit to the data, and make better predictions of the MCAS, indicating the model is useful, even given the flaws that might exist in it.

4 Conclusion & Future Work

It appears that we have found evidence that using students' responses to scaffolding questions were helpful in constructing a model that can track students' knowledge better (RQ1). Also, we presented results showing fine-grained models can better predict MCAS scores (RQ2). The important results presented are certainly about RQ2, where we show one instance where a fine-grained model can be used to predict students' skills better[17].

We believe that the ASSISTment system can be a better predictor of state test scores because of this work. Of course, teachers want reports by skills, and this is the first evidence we have saying that our skill mappings are "good". (We make no claim that our WPI-78 is an optimal fitting mapping.) We have shared our data with other scholars. Researchers interested are welcomed to contact us for detail.

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, careless slip and lucky guesses, learning and forgetting (Katz, Lesgold, Eggan & Gordin, 1992), 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. However, the expert-built models are subject to the risk of "expert blind spot", an education phenomenon documented in numerous prior studies (Koedinger, Alibali, & Nathan, 2000; Koedinger & Nathan, 1997, 2004; M. J. Nathan & Koedinger, 2000, 2003; M. J. Nathan, Koedinger, & Alibali, 2001; M. J. Nathan, Long, & Alibali, 2002; Mitchell J. Nathan & Petrosino, 2003; M. J. Nathan, Stephens, Masarik, Alibali, & Koedinger, 2002). We are happy to see that our first cognitive model fits well on student performance data. On top of that we are now engaged in an effort to refine the current model by bringing in content specialists, cognitive scientists and researchers in the field of student modelling. The plan is 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.

[17] Pardos, Heffernan, Anderson & Heffernan (2006) simultaneously worked to answer the same research question, using Bayesian networks, and they reached the same conclusion as we did (Pardos, Feng, Heffernan & Heffernan, 2007).

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 and 2005-2006 WPI/CMU ASSISTment Teams including Nathaniel O. Anozie, Elizabeth Ayers, Brian Junker, Andrea Knight, Meghan Myers, Carolyn Rose all at CMU, Steven Ritter at Carnegie Learning, Tom Livak, Abraao Lourenco, Michael Macasek, Goss Nuzzo-Jones, Kai Rasmussen, Leena Razzaq, Terrence Turner, Ruta Upalekar, and Jason Walonoski all at WPI.

5 References

Anderson, J. R. & Lebiere, C. (1998). The Atomic Components of Thought. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anozie N. & Junker, B. (2006). Predicting End-of-year Accountability Assessment Scores from Monthly Student Records in an Online Tutoring System. Workshop on Educational Data Mining held at the 21st National Conference on Artificial Intelligence (AAAI), Boston, 2006.

Ayers, E. & Junker, B. (2006). Do Skills Combine Additively to Predict Task Difficulty in Eighth-grade Mathematics? Workshop on Educational Data Mining held at the 21st National Conference on Artificial Intelligence (AAAI), Boston, 2006.

Barnes, T., (2005). Q-matrix Method: Mining Student Response Data for Knowledge. In Beck. J (Eds). *Educational Data Mining: Papers from the 2005 AAAI Workshop*. Technical Report WS-05-02. ISBN 978-1-57735-238-9.

Bock, R.D., Gibbons, R., & Muraki, E. J. (1988). Full Information Item Factor Analysis. Applied Psychological Measurement, 12, 261-280.

Cen. H., Koedinger K., & Junker B. (2005). Automating Cognitive Model Improvement by A*Search and Logistic Regression. In Beck. J (Eds). *Educational Data Mining: Papers from the 2005 AAAI Workshop*. Technical Report WS-05-02. ISBN 978-1-57735-238-9.

Corbett, A. T., Anderson, J. R., & O'Brien, A. T. (1995). Student Modeling in the ACT Programming Tutor. In P. Nichols, S. Chipman, & R. Brennan (Eds.). *Cognitively Diagnostic Assessment.* Hillsdale, NJ: Erlbaum.

Draney, K. L., Pirolli, P., & Wilson, M. (1995). A Measurement Model for a Complex Cognitive Skill. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Erlbaum.

Embretson, S. E. (1992) Structured Rasch models for measuring individual-difference in learning and change. *International Journal of Psychology* 27(3-4):372-372.

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Feng, M., Heffernan, N.T., & Koedinger, K.R. (2006a). Addressing the Testing Challenge with a Web-Based E-Assessment System that Tutors as it Assesses. *Proceedings of the Fifteenth International World Wide Web Conference*. pp. 307-316. New York, NY: ACM Press.

Feng, M., Heffernan, N.T, Koedinger, K.R. (2006b). Predicting State Test Scores Better with Intelligent Tutoring Systems: Developing Metrics to Measure Assistance Required. In Ikeda, Ashley & Chan (Eds.). *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Springer-Verlag: Berlin. pp. 31-40. 2006.

Feng, M., Heffernan, N.T, Murali 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.

Feng, M. & Heffernan, N. (2007). Towards Live Informing and Automatic Analyzing of Student Learning: Reporting in ASSISTment System. *Journal of Interactive Learning Research*. 18 (2), pp. 207-230. Chesapeake, VA: AACE.

Feng, M., Beck, J,. Heffernan, N. & Koedinger, K. (2008). Can an Intelligent Tutoring System Predict Math Proficiency as Well as a Standardized Test? In Baker & Beck (Eds.). *Proceedings of the 1st International Conference on Education Data Mining*. pp.107-116. Montréal 2008.

Feng, M., Heffernan, N., Beck, J, & Koedinger, K. (2008). Can we predict which groups of questions students will learn from? In Baker & Beck (Eds.). *Proceedings of the 1st International Conference on Education Data Mining*. pp.218-225. Montréal 2008.

Fischer, G.H. & Molenaar, I.W. (1995). *Rasch Models: Foundations, Recent Developments, and Applications*. New York: Springer Verlag.

Fischer, G. & Seliger, E. (1997). Multidimensional Linear Logistic Models for Change. Chapter 19 in van der Linden, W. J. and Hambleton, R. K. (eds.) (1997). *Handbook of modern item response theory*. New York: Springer Verlag.

Gierl, M.J., Wang, C., & Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in Algebra on the SAT. *Journal of Technology, Learning, and Assessment*, 6(6). Retrieved May, 2008 from http://www.jtla.org.

Hedeker, D. & Gibbons, R. D. (in progress). "Longitudinal Data Analysis": "Mixed-Effects Regression Models for Binary Outcomes" (chapter 9).

Heffernan N.T., Turner T.E., Lourenco A.L.N., Macasek M.A., Nuzzo-Jones G., Koedinger K.R. (2006). The ASSISTment Builder: Towards an Analysis of Cost Effectiveness of ITS creation. In Sutcliffe, G. & Goebel, R.(eds.). *Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference (*FLAIRS2006). pp. 515-520.

Junker, B. (2007). Using on-line tutoring records to predict end-of-year exam scores: experience with the ASSISTments project and MCAS 8th grade mathematics. To appear in Lissitz, R. W. (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard settings*. Maple Grove, MN: JAM Press.

José C. P. & Douglas M. B. (2000). *Mixed-Effects Models in S and S-PLUS*, Statistics and Computing Series, Springer-Verlag, New York, NY, 2000.

Mendicino, M., Heffernan, N. T. & *Razzaq, L.* (In Press) Comparison of Traditional Homework with Computer Supported Homework. *Journal of Research on Technology in Education (JRTE).* Published by the International Society For Technology in Education (ISTE).. Scheduled to appear in the Winter 2008 issue.

Mislevy, R. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.) *Educational measurement* (4th ed., pp.257-306). Washington, DC: American Council on Education.

Nichols, P. D., Chipman, S. F., & Brennan (1995) (Eds.). Cognitively Diagnostic Assessment. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Pardos, Z. A., Heffernan, N. T., Anderson, B., & Heffernan C. (2006). Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks. Workshop on Educational Data Mining held at the 8th International Conference on Intelligent Tutoring Systems, Taiwan, 2006.

Pardos, Z., Feng, M. & Heffernan, N. T. & Heffernan-Lindquist, C. (2007). Analyzing fine-grained skill models using bayesian and mixed effect methods. In Luckin & Koedinger (Eds.) In *Proceedings of the 13th Conference on Artificial Intelligence in Education*.pp.626-628. IOS Press: Amsterdam, Netherlands.

R Development Core Team (20070. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.r-project.org.

Raftery, A. E. (1995). Bayesian model selection in social research. In Sociological Methodology, 25, 111-163.

Razzaq, L., Heffernan, N. T. (2006). Scaffolding vs. Hints in the Assistment System. In Ikeda, Ashley & Chan (Eds.). *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Berlin: Springer-Verlag. pp. 635-644. 2006.

Razzaq, L., Heffernan, N. T., Lindeman, R. W. (2007) What level of tutor interaction is best? In Luckin & Koedinger (Eds) Proceedings of the 13th Conference on Artificial Intelligence in Education. IOS Press. pp. 222-229.

Singer, J. D. & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Occurrence*. Oxford University Press, New York.

Snijders, Tom A. B., and Bosker, Roel J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, London etc.: Sage Publishers, 1999.

Tan, E. S., Imbos, T. & Does R. J. M. (1994) A distribution-free approach to comparing growth of knowledge. *Journal of Education Measurement,* 31 (1):51-65.

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.

van der Linden, W. J. and Hambleton, R. K. (eds.) (1997) *Handbook of Modern Item Response Theory*. New York: Springer Verlag.

Yun, J. T., Willet. J. & Murnane, R. (2004) Accountability-Based Reforms and Instruction: Testing Curricular Alignment for Instruction Using the Massachusetts Comprehensive Assessment System. Paper presentation at the Annual American Educational Research Association Meeting. San Diego, 2004. Achieved at http://nth.wpi.edu/AERAEdEval2004.doc