

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

Mingyu Feng¹, Neil Heffernan¹, Murali Mani¹, Cristina Heffernan²

¹Worcester Polytechnic Institute, {mfeng|nth|mmani}@wpi.edu

²Worcester Public Schools, cristina_heffernan@charter.net

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 “skill”. Can we predict state tests scores better if we tag the questions with fine-grained models the skills needed? Psychometricians don’t do this presumably because they don’t get better fitting model, for a variety of reasons. We are investigating if we can do better prediction with finer-grained skill models. Our result gave a confirmative answer to this question.

Introduction

Most large standardized tests (like the math-subtest 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 knowledge component (i.e., 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 prerequisites relationship between these skills. Another reason is that students usually do a small number of items in a given setting (39 items for the 8th grade math Massachusetts Comprehensive Assessment System test). 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. We consider 3 different *skill models*¹, one that is unidimensional, which we will call the “WPI-1”, one that has 5 skills we call the “WPI-5”, and our most fine-grained

model has 78 skills we call the “WPI-78”. In all cases, a skill model is a matrix that relates questions to the skills needed to solve the problem. The WPI-1, WPI-5, and WPI-78 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.

Given that the WPI-78 composed of 78 skills, people might think the WPI-78 would naturally fit the data better than the skill models that contain far less skills. Moreover, they may even worry about 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 ASSISTment data based on which the models will be constructed. Instead, we used totally different data (from the external, paper-and-pencil based state test) as the testing set. Hence, we argue that overfitting would not be a problem in our approach.

Modeling student responses 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 and Junker, 2006, in press) are engaged trying to allow multi-mapping² using a version of the WPI-78 but report their LLTM model does not fit well. Anozie & Junker (2006, in press), 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 in that 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

Copyright © 2006, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

¹ What we refer to as a “skill model” is referred to as “Q-Matrix” by some AI researchers (Barnes, 2005) and psychometricians (Tatsuoka, 1990); and in Hao, Koedinger & Junker (2005), they used the term “cognitive model”, while Croteau, Heffernan & Koedinger (2004) used the term “transfer model”.

² 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 skills.

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 we have known, little effort has been put in that compares of different grain-sized skill models in the intelligent tutoring system area. The few that have done this is in a non-multi-mapped manner (Corbett, Anderson, & O'Brien, 1995; Draney, Pirollo, & 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).

The Massachusetts Comprehensive Assessment System (MCAS)

MCAS is a Massachusetts state administered standardized test that produces tests for English, math, science and social studies for grades 3 to 10. We focused on only 8th grade mathematics. Our work is related to the MCAS in two ways. First we built out content based upon released items. Secondly, 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. The state test consists of 5 open response, 4 short answer and 30 multiple choice (out of 4) questions. Only the multiple choice and short answer questions are used in our prediction 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 data set, the mean score out of 34 points was 17.9 (standard deviation=7.1).

Background on the ASSISTment Project

The ASSISTment system is an online tutoring system that is about 2 years old. In the 2004-2005 school year some 600+ students used the system about every two weeks. 8 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. In Massachusetts, the state department of education has released 8 years (1998-2005) worth of MCAS test items, over 300 items, which we have turned into ASSISTments by adding "tutoring". If students got the item correct they were given a new one. If they got it wrong, they were provided with a small "tutoring" session where they were forced to answer a few questions that broke the problem down into steps. The key feature of ASSISTments is that they provide instructional assistance while assessing students. Razzaq & Heffernan (2006, in press) addressed student learning due to the instructional assistance, while

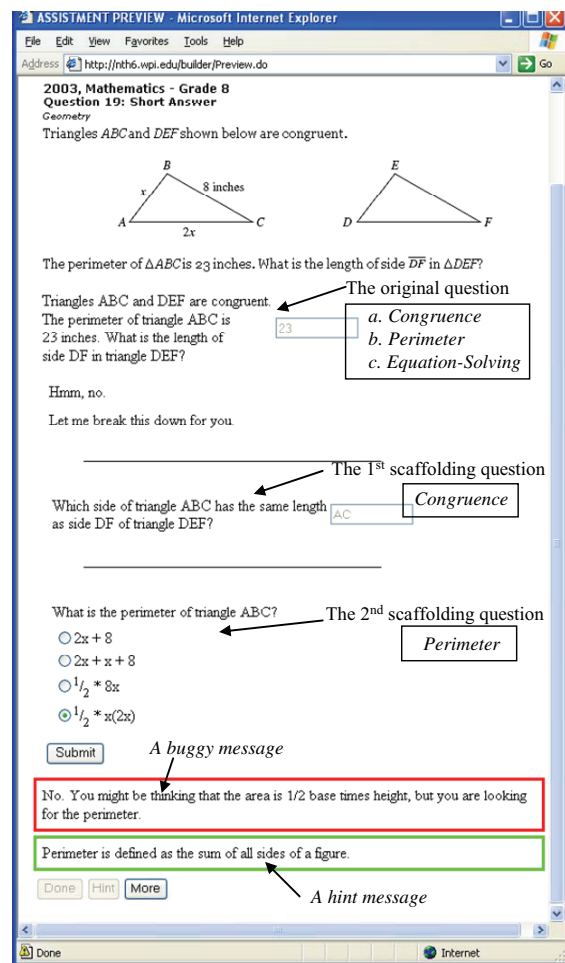


Figure 1. An ASSISTment shown just before a student hits the "Submit" button, showing 2 scaffolding questions, one buggy message and a hint message that can occur at different points.

this paper is focused on skill model evaluation by assessing students' performance on a state test.

Each ASSISTment consists of an original question and a list of scaffolding questions. The original question usually has the same text as in MCAS test while the scaffolding questions were created by our content experts to coach students who fail to answer the original question. An ASSISTment that was built for item 19 of the 2003 MCAS is shown in Figure 1. In particular, Figure 1 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 1 shows a buggy messages that appeared after the student clicked on “ $\frac{1}{2} * x(2x)$ ” 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

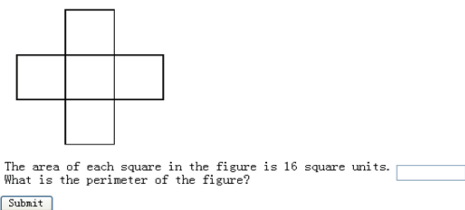


Figure 2. The original question of item 27 of 1999 MCAS test

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.

Figure 2 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”.

How was the Skill Models Created?

In April, 2005, we staged a 7 hour long “coding session”, where our subject-matter expert, Cristina Heffernan, with the assistance of the 2nd author set out to make up skills and tag all of the existing 8th grade MCAS items with these skills.³ There were about 300 released test item for us to code. Because we wanted to be able to track learning between items, we wanted to come up with a number of

³ We hand-coded the skills in this work. Though, we believe it is possible to use an automatic technique such as LFA (Hao, Koedinger & Junker, 2005) or Q-matrices (Barnes, 2005) for topic construction.

skills that were somewhat fine-grained but not too fine-grained such that each item had a different skill. We therefore imposed upon our subject-matter expert that no one item would be tagged with more than 3 skills. She was free to make up whatever skills she thought appropriate. We printed 3 copies of each item so that each item could show up in different piles, where each pile represented a skill. Although we have English names for the skills, those names are just a handy tag; the real meaning of a skill must be divined by 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 6 foot-long tables covered with 106 piles of items. 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-78⁴. To create the coarse-grained models, such as the WPI-5, we used the fine-grained model to guide us. We decided to use the same 5 categories that both the National Council of Teachers of Mathematics uses, as well as the Massachusetts Department of Education. These categories 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 categories, but our mapping was not the same as the states’. 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. 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 and WPI-78.

The WPI-1 and WPI-5 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-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 the item 19 from 2003 MCAS test (See Figure 1). 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-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 the item 27 from 1999 MCAS test shown in Figure 2 is tagged “Perimeter” and “Area”, and therefore it is tagged with “Measurement” in the WPI-5, and then again “math” in the WPI-1.

⁴ In Pardos, Heffernan, Anderson & Heffernan (2006, in press) we called this model the WPI-106 because they used a data set that included additional items.

Table 1. Hierarchical relationship among skill models

	WPI-78	WPI-5	WPI-1
	Inequality-Solving	Patterns, Relations and Algebra	The skill of "Math"
	Equation-Solving		
	Equation-Concept		
	...		
	Plot Graph		
	X-Y-Graph	...	
	...	Geometry	
	Congruence		
	...		
	Similar-Triangles	Measurement	
	Perimeter		
	Circumference		
	...		
	Area		

Research Questions

Specifically, the research questions we try to answer can be formulated as below:

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?

We think that that getting an answer to RQ1 would help us properly evaluate the second and more important research question.

Research Question 2 (We will refer to as RQ2): How does the finer-grained skill model (WPI-78) do on estimating external test scores comparing to the skill model with only 5 categories (WPI-5) and the one even with only one category (WPI-1)?

We think that an answer to RQ2 that said that finer-grained models allow for better modeling/prediction would have important societal implications (e.g. regarding to tracking student performance.)

Comparing the capability of student knowledge tracking of different grain-sized skill models is the main goal of this work. Though, since we were using the accuracy of the predicted MCAS test score as the gauge of model performance, we extended our goal a little bit towards building predictive models and brought up the third research question as below.

Research Question 3 (We will refer to as RQ3): Does introducing item difficulty information help to build a better predictive model and thus more precisely estimate students’ score on the external test?

We observed that items tagged with similar skills vary on their difficulty due to different reasons such as the context, is the question asked directly or indirectly or what are the numbers involved in the problem. Therefore, we hypothesize that using item difficulty as a covariate may help us to improve our model and thus predict students’ performance on the external test more accurately.

Approach

Data Source

We collected data of 497⁵ students who used our system from Sep. 17, 2004 through May 16, 2005 for on average 7.3 days (one period per day)⁶. All these students have worked on the system for at least 6 days. We excluded data from the students’ first day of using the system considering the fact they were learning how to use the system at that time. The item-level state test report was available for all these 497 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 and 189 scaffolding questions.

The data is organized in the way that there can be one or multiple rows for every student response to each single question depending on what’s the 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

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 May 17, 2005, it would be inappropriate to use data after that day for the purpose of predicting state scores. Therefore those 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

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 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’s response on item 19 (with original item ID⁷ being 326) and the rest 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 row 1 - 3 and so does the original question of item 27 as in row 9 and row 10. For both items, the student answered the original questions wrong (indicated by “0” in the response column of row 1-3 and row 9-10) and thus was presented 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 the other scaffolding questions. On contrast, for item 27, though not getting the original question right on the first shot, the student went through all three scaffolding questions correctly. WPI-78 is the skill model being used here.

The Statistical Model we Fit - Mixed-effects Logistic Regression Model

For dichotomous (binary in our case) response data, several approaches adopting either a logistic or probit regression model and various methods for incorporating

and estimating the influence of the random effects have been developed. Snijders & Bosker (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 on 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 $x_i = (1, x_{i1}, x_{i2}, \dots, x_{ip})$ denote the set of covariates and $\beta = (\beta_0, \beta_1, \dots, \beta_p)'$ 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:

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

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

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:

⁷ 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.

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

Level-2 (or between-person) model:

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

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

Where

p_{ij} is the probability 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” and tells how good is the student’s initial knowledge; b_{1i} represents the “slope” and tells what’s 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 “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 R⁸, 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

```
## 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)
```

ran in R to train a mixed-effects logistic regression model using both TIME and WPI-5 skills as covariates.

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.

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

⁸ <http://www.r-project.org/>

of the 3 skill models by our subject matter expert. 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 item-score?

The first thing we did is identifying what are the skills associated with the item in that skill model. Thus, given a student’s learning parameters on related skills (and the item difficulty level of each item in the state test, depending which model was used), for any particular item in the state test, we can calculate the probability of positive response from the student. In the case that an item was tagged with more than one skill (i.e., when WPI-5 and WPI-78 was used as the skill model), we picked the lowest probability among all the skills that apply to the item⁹ for that student¹⁰. 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. Therefore, once we acquired the probability of correct response for all the items, we sum them up to produce the total points awarded. For example, if the probability of an item marked with Geometry is 0.6, then 0.6 points are added to the sum to produce the points awarded. This sum of these points is what we use as our prediction of their state test score¹¹.

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 Mean Absolute Difference (**MAD**). We also calculate a normalized metric named **% Error** by dividing the MAD by 34 to reflect the fact that the MAD is out of a full score 34 points.

Does Adding Scaffolding Questions Help?

Scaffolding Credit and Partial Blame

We started our work examining only students’ response to original questions. And then we brought up RQ1, asking ourselves that if we can improve our models by including students’ response to the scaffolding questions. As

⁹ 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 colleague Pardos, Hefernan, Anderson and Heffernan (2006, in press) use this approach and seem to getting similar results that fine grained models enable better predictive models.

¹⁰ We consider the skill that had the lowest probability of correct response in our model the hardest skill for a student.

¹¹ We think it might be useful to discuss your 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 not, in that sense that if you tagged a bunch of item with a easy skill (i.e., one easier than what the item was currently tagged with), that would not change our models 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.

Table 3. The effect of using scaffolding questions

Skill Model	MAD		% Error (MAD/34)		Δ %Error	p-value of paired t-test
	Orig. Response	Orig.+ Scaffolding Response	Orig. Response	Orig.+ Scaffolding Response		
WPI-1	5.269	4.552	15.50%	13.39%	2.11%	0.0043
WPI-5	5.671	4.343	16.68%	12.77%	3.91%	<0.0001
WPI-78	6.035	4.121	17.75%	12.12%	5.64%	<0.0001

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, if a student gets an original question wrong, his/her 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 incorrectly an item tagged with multiple skills. 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 adopted the “partial blame” strategy: if a student got such an item wrong, the skills in that item will be sorted according the overall performance of that student on those skills and only the one 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.

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 achieved by using scaffolding questions. In particular when using the WPI-1, the mean decrease of “% Error” is 2.1% after scaffolding questions were introduced; for WPI-5, the number is 3.9%;

and the number is 5.6% for WPI-78. We then did paired-t-tests between the “% Error” terms for the 497 students and found that the improvements are statistically significant in all the three cases as summarized in Table 3. [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.

Does an error rate of 12.12% on the WPI-78 seem impressive or poor? What is a reasonable goal to shoot for? Zero percent error? In Feng, Heffernan & Koedinger (2006, in press) we reported on a simulation of giving two MCAS 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.

Does the Finer Grained Model Predict Better?

Does WPI-78 Fit Better than WPI-5? How about WPI-1?

To answer RQ2, we compared the three mixed-effects regression models (trained on the “full” data set with scaffolding questions used) fitted using the 3 different skill models. As shown in Table 4 (most content extracted from Table 3), the WPI-78 had the best result, followed by the 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-5 to WPI-78.

Table 4. Evaluating the accuracy of skill models

Skill Model	MAD	95% Confidence Interval for MAD	% Error (MAD/34)
WPI-1	4.552	[4.256, 4.848]	13.39%
WPI-5	4.343	[4.066, 4.620]	12.77%
WPI-78	4.121	[3.860, 4.382]	12.12%

To see if these “% Error” were statistically significant different for the models, we compared each model with every other model. We did paired-t-tests between the “% Error” terms for the 497 students and found that the WPI-78 model is statistically significantly better than the WPI-5 and WPI-1 ($p < .0001$ in both cases), and WPI-1 is statistically significantly worse than the other two models ($p < .0001$ in both cases). This suggested that using finer-grained skill models was helpful on 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 because of the result presented in Tables 3 & 4. Therefore, though questions such as “Are the improvements in accuracy from 4.552 to 4.121 meaningful?”, “What is the practical value of this improvement?” are interesting, they are beyond the scope of this paper.

Readers may have noticed that when only response data on original questions were used, the order reversed: WPI-78 was doing worse than WPI-5 and WPI-5 led to a higher MAD than WPI-1. 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 model 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 them might get. The students in our data set completed, on average, over 270 questions.

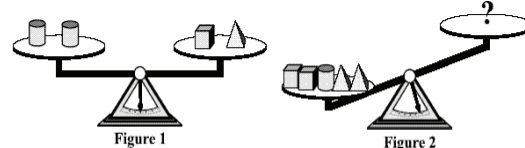
How well does the Model WPI-78 Fit the Data?

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 a model we have. 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 student knowing them better, while the slope allows for the fact that some skills are learned more quickly than others. In our model, the easiest skill was “Ordering-Numbers” (e.g., “Which of the two numbers is larger? -2 or -1”), while the skill that had the hardest incoming difficulty was “Pythagorean-theorem”, a skill that 8th graders in our district are expected to know at the beginning of the year. We also looked at the fits on the slopes for each skill. The skill that showed the steepest rate of learning during the course for 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 that showed the least learning was called “Equation-Concept” (as shown in Figure 3). 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”). This refinement *might* work better if students’ learning of equation-concept should transfer to increase their performance on equation-solving and vice-versa.

We did a correlation to see if the skills that were poorly fit for their slopes were the same skills that had a relatively smaller numbers of items, but surprisingly we found that the correlation was very weak. In addition, we speculated that skills that had less data for them would be more likely to be poorly fit. To investigate that we sorted the 78 items by the amount of data we had for each, and found that 8 of the worse fitting skills (as identify above by high p-value) also had a small amount of data. Specifically, these 8 skills had less than 440 data instances, while the median amount of data for each skill was 1391 records. This confirmed our expectation that skills with less data will be less well fit. Out of the 78 skills, 13 coefficients predicted un-learning (i.e. the slopes are negative), and 10 of those coefficients were not statically significant different from zero. Overall, 40 of the slope-coefficients were not statically significantly different than zero. Again, in the future we will automate the process to remove those skills from the model and to re-fit the data. 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.

We further looked to see which skills were both poorly fit for their intercept (15) and slopes (40). It turned out that there were 11 instances of skills that were poorly fit along both lines. These are instances of skills that need further refining in our model.



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

Figure 3. A question tagged with the skill “Equation-Concept”

Does Introducing Item Difficulty Help?

Getting Item Difficult Parameters

In order to answer RQ3, item difficulty information was required for the ASSISTments as well as the state test items. To obtain such information, we adopted Item Response Theory (IRT) and trained Rasch models

(Embretson & Reise, 2000; van der Linden & Hambleton, 1997), based on online data and the state test responses.

IRT now underlies several major tests. Computerized adaptive testing (CAT), in particular, relies on IRT. In CAT, examinees receive items that are optimally selected to measure their potential. IRT principles are involved in both selecting the most appropriate items and equating scores across different subsets of items. IRT now contains a large family of models. The simplest model is the Rasch model, also known as the one-parameter logistic model (1PL). For the model, the dependent variable is the dichotomous response for a particular person to a specified item. The independent variables are the person's trait score, θ_s , and the item's difficulty level, β_i .

Though the Rasch model can be used to estimate the probability of the success response on specified items, it is a unidimensional model. Therefore, our approach is to obtain the β from the trained Rasch models for all items (discarding θ) and use β as a covariate in our mixed-effects logistic regression models to incorporate item difficulty with different grain-sized skills models.

To get the β 's for the ASSISTments, we were using 2005-2006 ASSISTment data for the same group of items but done by a different group of 2702 students from the same district as the 497 students in our data, assuming students from different years are of the same knowledge level. After training up the Rasch model, we extracted the β 's for all the items and observed that the values of β center around zero and range¹² from -2.37 to 2.69. Then we added a new column in our data (See the sample data in Table 1.) putting in the corresponding β for the particular item in the each row. Now the data is ready to be used to train mixed-effects logistic regression models with β as a covariate. The similar approach was followed to get the β 's for the state test items. The item level response data of 1117 8th graders from Worcester who have not gotten involved in the ASSISTment system was utilized to train the Rasch model and we observed that the β 's of the 34 state test items range from -2.22 to 1.60.

Results and Discussion

With β acquired, we then updated the three regression models (with different skill models being used) as described in Section 4 by introducing β as a covariate to obtain 3 new models. Now we have three pairs of models with 2 models (without-beta and with-beta) for each skill model. To check if the internal validity of the introduction of β , we examined the Bayesian Information Criterion (BIC) of the three pairs of models and found that for each pair, there is a huge drop-down (over 10,000 points¹³) on BIC when β was used as a covariate, suggesting that by introducing β , we can construct statistically significantly

better fitted models on the ASSISTment data, tracking the change of students' performance. Afterwards, we got students' predicted score on the state test and calculated the MAD and % Error as we did before. Unfortunately, this time, no supportive evidence was found for the adding of item difficulty. The MAD and % Error are at about the same level for both models of all three pairs, suggesting that though item difficulty information was not as helpful to on estimating the score of the external state test as on tracking learning in the ASSISTment system. This result surprised us and we are still seeking plausible explanation for the result that an internally better fitted model does not serve as a better predictive model on the external measure.

Though failed to give a confirmative answer to RQ3, the work reassured the conclusion about RQ2. After β was introduced as a covariate, the relationship between the skill models remained the same: WPI-78 with-beta model is statistically significantly better and WPI-1 with-beta model is significantly worse than the other two models.

Conclusion

It appears that we have found evidence that using students' response to scaffolding questions were helpful to constructing a model that can track students' knowledge better (RQ1). Also, we presented results showing fine-grained models can better predict MCAS score (RQ2). Item difficulty was introduced as a factor to improve the predictive models and we ended up with better internally fitted models, but surprisingly no significant enhancement was observed on the prediction of the scores of the external state test (RQ3). Reasoning from a Null result is always dangerous. In fact we don't hypothesize that item difficulty will not help. If anything, our guess is that the 34 items from the MCAS might not be the most reliable way of accessing our models. The important results presented are certainly about RQ2, where we show one instance where fine-grained model can be used to predict students' skills better. Pardos, Heffernan, Anderson & Heffernan (2006, in press) simultaneously worked to answer the same research question, using Bayesian networks, and they reached the same conclusion as we did.

You will remember that for RQ2, we do not bother to report the internal fit of our models using measures like BIC, and there are two reasons. The first is that "internal fit is not the important variable" and the second is "we don't know how to". Less flippantly, the first reason is that what we really care about is the ability to predict the scores on the state test, so the internal fit of the model is not the main question of interest. Secondly, 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 skills, and BIC only make sense when the data is meant to be the exact same.

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

¹² A higher value of β indicates the item is relatively easy while a lower one suggests a relative harder item.

¹³ Raftery (1995) discussed a Bayesian model selection procedure, in which the author proposed the heuristic of a BIC difference of 10 is about the same as getting a p-value of $p = 0.05$.

“good”. (We make no claim that our WPI-78 is an optimal fitting mapping.) Now that we are getting reliable models showing the value of these models, we will consider using these models in selecting the next best-problem to present a student with. As part of the future work, we will get our data ready to be shared with other scholars. Researchers interested are welcomed to contact us for detail.

Acknowledgements

This research was made possible by the US Dept of Education, Institute of Education Science, “Effective Mathematics Education Research” program grant #R305K03140, the Office of Naval Research grant #N00014-03-1-0221, NSF CAREER award to Neil Heffernan, and the Spencer Foundation. All of the opinions in this article are those of the authors, and not those of any of the funders.

This work would not have been possible without the assistance of the 2004-2005 WPI/CMU ASSISTment Team 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.

References

- Anderson, J. R. & Lebiere, C. (1998). *The Atomic Components of Thought*. LEA.
- Anozie N. & Junker, B. (2006, in press). 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, in press). 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.
- Corbett, A. T., Anderson, J. R., & O'Brien, A. T. (1995). Student Modeling in the ACT Programming Tutor. Chapter 2 in P. Nichols, S. Chipman, & R. Brennan, *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, *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Erlbaum.
- Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, New Jersey.
- Feng, M., Heffernan, N.T, Koedinger, K.R., (2006, in press). Predicting State Test Scores Better with Intelligent Tutoring Systems: Developing Metrics to Measure Assistance Required, The 8th International Conference on Intelligent Tutoring System, 2006, Taiwan.
- Hao C., Koedinger K., and 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.
- Hedeker, D. & Gibbons, Robert. D. (in progress). “*Longitudinal Data Analysis*”: “Mixed-Effects Regression Models for Binary Outcomes” (chapter 9).
- José C. P. and Douglas M. B. (2000). Mixed-Effects Models in S and S-PLUS, *Statistics and Computing Series*, Springer-Verlag, New York, NY, 2000.
- Pardos, Z. A., Heffernan, N. T., Anderson, B., & Heffernan C. (2006, in press). 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.
- Raftery, A. E. (1995). Bayesian model selection in social research. In *Sociological Methodology*, 25, 111-163.
- Razzaq, L., Heffernan, N. T. (2006, in press). 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.
- 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.
- 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>