

# Investigating the utility of a conjunctive model in Q-matrix assessment using monthly student records in an online tutoring system\*

Nathaniel O. Anozie and Brian W. Junker  
Department of Statistics  
Carnegie Mellon University

Paper Prepared for the Annual Meeting of the  
National Council on Measurement in Education (NCME)  
Chicago, Illinois, USA  
April 10-12, 2007

March 22, 2007

---

\*This work would not have been possible without the efforts and insights of our many colleagues, including principal investigators Neil Heffernan (Worcester Polytechnic Institute) and Ken Koedinger (Carnegie Mellon) as well as Elizabeth Ayers, Andrea Knight, Meghan Myers, Carolyn Rose all at CMU, Steven Ritter at Carnegie Learning, Mingyu Feng, Tom Livak, Abraao Lourenco, Michael Macasek, Goss Nuzzo-Jones, Kai Rasmussen, Leena Razzaq, Terrence Turner, Ruta Upalekar, and Jason Walonoski all at WPI; and was supported with funding from the US Department of Education, National Science Foundation (NSF), Office of Naval Research, Spencer Foundation, and the US Army.

## Abstract

Heffernan et al. (2001) have developed an online intelligent tutoring system (the ASSISTment system) for eighth grade mathematics that is explicitly aligned to state exam standards. In this paper we assess the utility of the DINA model (Junker and Sijtsma, 2001) for diagnosing knowledge components (KC's: skills, pieces of knowledge, and other cognitive attributes) that students do and do not have, based on their interactions with the tutoring system. In a split-half cross-validation experiment we found that with 50 or more questions per student we could obtain 70% prediction accuracy. We also found that many students had attained 30% or less of the KC's associated with these questions in a Q-matrix based transfer model. As expected, the KC's students lack are not basic whole-number, addition, and similar skills, but higher-order symbolic skills such as algebraic manipulation and equation solving. Overall, we find that the DINA model is quite promising for on-line KC-based diagnosis and reporting.

keywords: diagnostic assessment, benchmark assessment, Bayes Net, intelligent tutor

## 1 Introduction

Recently there has been intense interest in using periodic benchmark tests to predict student performance on end-of-year accountability assessments (Olson, 2005). Benchmark tests are typically paper-and-pencil tests given at regular intervals, from three times a year to monthly, in order to predict progress toward proficiency on state accountability exams. Some benchmark tests also try to function as formative assessments for teachers, so that the classroom time occupied by the benchmark test is not completely lost to teachers' instructional mission. Nevertheless, teachers may still find the time needed for benchmark tests to be an intrusion on instructional time.

An alternative approach may be available when an online, computer-based tutoring system is in place. The benefits of online tutoring systems are well known: Koedinger, Corbett, Ritter & Shapiro (2000) study classroom evaluations of the Cognitive Tutor Algebra course (e.g.,

**19** Triangles  $ABC$  and  $DEF$  shown below are congruent.

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

Figure 1: A released MCAS item. This item would be rendered in similar format as a “main question” for one ASSISTment item.

Koedinger, Anderson, Hadley, & Mark, 1997; Koedinger, et al., 2000) and demonstrate that students in tutor classes outperform students in control classes by 50–100% on targeted real-world problem-solving skills and by 10–25% on standardized tests. Beck, Peng & Mostow (2003) argue for a mixed predictive/formative use for benchmarks based on tutor interaction logs for their online reading tutor.

Heffernan, et al. (2001) have developed an online tutoring system for eighth grade mathematics that is explicitly aligned to state exam standards, and in fact takes released state exam questions and morphs<sup>1</sup> of them as the main student tasks for tutoring. Their system is called the ASSISTment<sup>2</sup> system, since it implicitly assesses progress toward proficiency on the state exam at the same time it assists student learning in mathematics.

A typical student interaction in the ASSISTment System is built around a single released MCAS item, or a morph of a released item, from the end of year accountability exam (the 8th grade MCAS mathematics exam), for example as shown in Figure 1. The item would be rendered in the ASSISTment system in a similar format. This is called a “main question”. Figure 2 gives

<sup>1</sup>In other contexts, e.g. Embretson (1999), item morphs are called “item clones”.

<sup>2</sup>Coined by Ken Koedinger, to combine the *assisting* and *assessment* functions of the system.

What is the length of side DF in triangle DEF?

Original question

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

AB  BC  AC

Hint

What is the perimeter of triangle ABC?

$1/2 * 8x$    $2x + 8$    $2x + x + 8$    $1/2 * x(2x)$

Buggy message that shows up if the student selected  $1/2 * 8x$ . "No. You might be thinking that the area is  $1/2$  base times height, but you are looking for the perimeter."

Now, given the perimeter of triangle ABC equals 23 inches, you can write the equation  $2x + x + 8 = 23$  and solve it for  $x$ . What is the value of  $x$ :

Hint

Good. You've just got the value of  $x$ . Now you can get the length of side AC. What is it?

Scaffolding questions

Remember, we are looking for side DF. Enter the length of side DF:

Messages

Corresponding sides are congruent. In the picture below, corresponding sides are colored.

AC is equal to  $2x$ :

$AC = 2x$        $x = 5$   
 $AC = 2 * 5$   
 $AC = 10$

Figure 2: Annotated student interaction with the ASSISTment system, based on the main question in Figure 1.

an annotated view of the interaction that a student might have, based on the main question in Figure 1. If the student correctly answers the main question, a new main question is presented. If the student incorrectly answers, a series of “scaffolding” questions are presented, breaking the main question down into smaller, learnable chunks. The student may request hints at any time, and if the student answers a question incorrectly, a “buggy message” keyed to a hypothesized bug or error in the student’s thinking is presented. Multiple hints on the same question become increasingly specific. The student repeatedly attempts each question until correct, and then moves on to the next question. Each package of a main question and its associated scaffolds is a single ASSISTment item.

Two statistical goals for the ASSISTment system are to predict end-of-year MCAS scores, and to provide regular, periodic feedback to teachers on how students are doing, what to teach next, etc. These goals are complicated in several ways by ASSISTment system design decisions that serve other purposes. For prediction, the exact content of the MCAS exam is not known until several months after it is given, and ASSISTments themselves are ongoing throughout the school year as students learn (from teachers, from ASSISTment interactions, etc.). Junker (2006) reviews

progress on the prediction goal of the ASSISTment project, using methods ranging from linear regression (Anozie & Junker, 2006) and HLM-based growth curve models (Feng et al., 2006) to item response theory (Ayers & Junker, 2006) and Bayes net models (Pardos et al., 2006).

In this paper we focus on the goal of diagnosing how students are doing: what knowledge components (KC's; e.g. skills, pieces of knowledge, and other cognitive attributes) students have obtained, and what they still need to learn. This goal is also complicated by practical features of the system. Different transfer models<sup>3</sup> are used and expected by different stakeholders: the MCAS exam itself is scaled using a unidimensional item response theory (IRT) model (van der Linden & Hambleton, 1997), but description and design of the MCAS is based on a five-strand model of mathematics (Number & Operations, Algebra, Geometry, Measurement, Data Analysis & Probability) and 39 “learning standards” nested within the five strands. In addition, ASSISTment researchers who have examined MCAS questions have developed a transfer model involving up to 106 KC's (WPI-106, Pardos et al., 2006) nested within the 39 learning standards, some 77 of which are active in the ASSISTment content considered in the present work. To the extent possible, feedback reports should be delivered at the granularity expected by each stakeholder. Moreover, scaffolding questions have an ambiguous status in practice: they can be designed as measures of single KC's in a particular transfer model, thus improving measurement of those KC's; or they can be designed to be optimal tutoring aids, regardless of whether they provide information on particular KC's in a particular transfer model<sup>4</sup>. Finally, different students work through ASSISTments at different rates, depending on their own work pace, attendance records, and the curriculum being followed by their classroom teachers; consequently the “experimental design” and sample size for diagnosing particular KC's varies from student to student.

In this paper we report on an initial use of the DINA (Deterministic Inputs, Noisy AND-gate; Sijtsma & Junker, 2001) conjunctive Bayes net model for cognitive diagnosis with the WPI-106

---

<sup>3</sup>A *transfer model* specifies the KC's needed to solve a problem, and might be coded with a Q-matrix, as in Embretson (1984), Tatsuoka (1990) or Barnes (2005).

<sup>4</sup>It can be argued that good tutorial scaffolds do focus on single KC's or small sets of KC's in *some* relevant transfer model, but in a multiple-transfer-model environment, scaffold questions need not map well onto KC's in *every* relevant transfer model.

transfer model. In Section 2 we briefly describe our data as well as some imputation methods that are needed for our work. In Section 3 we describe the DINA model and MCMC estimation methods for it. In Section 4 we describe some results of our investigations: split-half accuracy rates for predicting correctness of answers to cross-validation questions, tracking overall, student-level and KC-level mastery over time, and examining DINA parameter estimates for insights into the functioning of questions and the transfer model. Finally in Section 5 we discuss some implications and next steps for our work.

## 2 Data

The data we consider comes from from the 2004-2005 school year, the first full school year in which ASSISTments were used in classes in two middle schools in the Worcester School district in Massachusetts. At that time, the ASSISTment system contained a total of 493 main questions and 1216 scaffolds; 912 unique students logs were maintained in the system over the time period from September to April. Of these, approximately 400 main questions and their corresponding scaffolds were in regular use. The remaining questions and students represented various experimental or otherwise non-usable data for the studies considered here. We consider six months of tutor system data, October 2004 through March 2005. The number of students in each month ranged from 400 to 600, the number of KC's studied in each month ranged from 50 to 70, and the number of questions items studied in each month ranged from 400 to 1200.

Although the system is web-based and hence accessible in principle anywhere/anytime, students typically interact with the system during one class period in the schools' computer labs every two weeks. Because ASSISTment items were assigned randomly to students within curricula developed by teachers and researchers, and because students spent varying amounts of time on the system, the sample of ASSISTment items seen by each student varied widely from student to student. We treat the items that students did *not* see for these reasons as *missing completely at random* (MCAR; Mislevy and Wu, 1996); terms corresponding to MCAR data are simply dropped from the likelihood in the formal modeling below.

Another kind of missing data occurs because, as discussed in Section 1, the ASSISTment system employs a forced scaffolding strategy: if a main question is answered incorrectly, students must complete a sequence of related scaffolding questions; but if a main question is answered correctly, the corresponding scaffold questions are skipped. Missing scaffold question responses are therefore very informative—such responses are missing only if the main question was gotten right. This is an example of data *not missing at random* (NMAR; Mislevy and Wu, 1996). In order to capture as much information as possible in estimating whether or not students possess particular KC’s, we decided to impute these missing responses by crediting all scaffold questions corresponding to each correctly-answered main question.

Underlying this imputation scheme is the assumption that KC’s for the scaffold questions are necessary for correctly answering the main question. This seems like a safe assumption when the main question is tagged with several KC’s in the transfer model and each scaffold is designed to measure one of those KC’s. It is a less satisfactory assumption when scaffolds are designed to be optimal tutoring aids, or when they were designed with a different transfer model in mind, since in that case they may involve KC’s not employed in the main question (e.g. if the main question were designed to tap a cross-multiplication skill for proportional reasoning, and a corresponding scaffold question tapped a coordinated count-up skill instead).

### 3 Modeling and Estimation

Let  $i = 1 \dots I$  denote students,  $j = 1 \dots J$  denote questions, and  $k = 1 \dots K$  denote KC’s. The basic ingredients of many cognitive diagnosis models (Junker, 1999) are the item response variables

$X_{ij} = 1$  or  $0$ , indicating whether student  $i$  answered question  $j$  correctly;

the elements of the transfer model or  $Q$ -matrix

$Q_{jk} = 1$  or  $0$ , indicating whether knowledge component  $k$  is relevant to question  $j$ ;

and the indicators of KC’s for individual students

$\alpha_{ik} = 1$  or  $0$ , indicating whether student  $i$  possesses knowledge component  $k$ .

### 3.1 The DINA Model

The DINA model (so named by Junker & Sijtsma, 2001) is a simple conjunctive Bayes net model for cognitive diagnosis. It has been used as the basis of many approaches to cognitive diagnosis and assessment (e.g. Macready & Dayton, 1977; Haertel, 1989; Tatsuoka, 1990). In the DINA model, latent response variables  $\xi_{ij}$  are defined as

$$\xi_{ij} = \prod_{k: Q_{jk}=1} \alpha_{ik} = \prod_{k=1}^K \alpha_{ik}^{Q_{jk}}$$

indicating whether examinee  $i$  has all the KC's required for question  $j$ . In Tatsuoka's (1990) terminology the latent vectors  $(\alpha_{i1}, \dots, \alpha_{iK})$  are called *knowledge states*, and the vectors  $(\xi_{i1}, \dots, \xi_{ij})$  are called *ideal response patterns*.

The latent response variables  $\xi_{ij}$  are related to the observable response variables  $X_{ij}$  with two parameters, the slip parameter  $s_j$  and the guessing parameter  $g_j$ :

$$P(X_{ij} = 1) = \begin{cases} 1 - s_j & \text{if } \xi_{ij} = 1 \\ g_j & \text{if } \xi_{ij} = 0 \end{cases}$$

The slip and guessing parameters are related to question difficulty: questions with large  $s_j$  and small  $g_j$  are more difficult than questions with small  $s_j$  and large  $g_j$ . Generally speaking a question with low  $s_j$  and low  $g_j$  is more discriminating—that is, there is a closer relationship between  $X_{ij}$  and  $\xi_{ij}$ —than questions with large  $s_j$  and large  $g_j$ . When both  $s_j$  and  $g_j$  are large, this may suggest either that the question should be rewritten to more clearly tap the relevant KC's, or the transfer model ( $Q$ -matrix) should be redesigned to associate a more relevant set of KC's to that question. Dibello, Stout & Roussos (1995) address similar issues in their discussion of the *positivity* of a question with respect to the KC's it measures.

Given the response  $X_{ij}$ , the DINA model affords a natural way to update the odds that a student possesses KC  $k$ , using Bayes' rule (Junker & Sijtsma, 2001, p. 267):

$$\text{Odds}_{\text{after seeing } X_{ij}}(\alpha_{ik} = 1) = c_{ijk} * \text{Odds}_{\text{before seeing } X_{ij}}(\alpha_{ik} = 1)$$

where

$$c_{ijk} = \begin{cases} \frac{s_j}{1-g_j} & \text{if student has all KC's, besides } k, \text{ for question } j, \text{ and } X_{ij} = 0 \\ \frac{1-s_j}{g_j} & \text{if student has all KC's, besides } k, \text{ for question } j, \text{ and } X_{ij} = 1; \\ 1 & \text{otherwise.} \end{cases}$$

More details are provided by Junker & Sijtsma (2001) and Junker (1999). de la Torre and Douglas (2004) provide a comparative evaluation of the DINA model with the LLTM (linear logistic latent trait model); the DINA model worked well even on data simulated by the LLTM model.

### 3.2 The Likelihood, Prior and Posterior

Applying the principle of local independence, as in item response theory models, we can easily see that the likelihood for the DINA model when all students answer all questions is

$$P(X|\alpha, s, g) = \prod_{i=1}^I \prod_{j: i \text{ sees } j} [s_j^{1-X_{ij}}(1-s_j)^{X_{ij}}]^{\xi_{ij}} [(1-g_j)^{1-X_{ij}}g_j^{X_{ij}}]^{1-\xi_{ij}} \quad (1)$$

where the notation “ $j: i \text{ sees } j$ ” in the inner product is intended to account for students working on different subsets of the ASSISTments depending on their work pace, computer lab availability, teachers’ selection of curricula, etc.: only terms corresponding to questions each student actually sees are incorporated into the inner product; we are treating this kind of missingness as missing completely at random. Missing scaffold question responses corresponding to correct main questions are imputed to be correct (see Section 2) and so do not affect Equation 1. We also assume that each KC has prior probability  $p_k \equiv P(\alpha_{ik} = 1)$  of being present in the relevant population of students as a whole.

Following de la Torre & Douglas (2004) we take all prior distributions to be 4-beta distributions<sup>5</sup>. In particular we make the following prior assumptions:

- $p_k \sim \text{beta}(1, 1)$  for all  $k$ ;
- $s_j \sim 4\text{-beta}(2, 2, 0, 0.5)$  for all  $j$ ;

---

<sup>5</sup>A random variable  $V$  has a 4-beta( $\alpha, \beta, a, b$ ) distribution if and only if it can be written as  $V = a + (b - a)B$ , where  $B$  has a beta( $\alpha, \beta$ ) distribution.

- $g_j \sim 4\text{-beta}(2, 2, 0, 0.5)$  for all  $j$ ;

It is then easy to calculate complete conditional (posterior) distributions for each parameter, given the data and the other parameters:

- $p_k \sim \text{beta}(\sum_{i=1}^I \alpha_{ik} + 1, I - \sum_{i=1}^I \alpha_{ik} + 1)$
- $g_j \sim 4\text{-beta}(\sum_{i=1}^I X_{ij}(1 - \xi_{ij}) + 2, \sum_{i=1}^I (1 - X_{ij})(1 - \xi_{ij}) + 2, a = 0, b = 0.5)$
- $s_j \sim 4\text{-beta}(\sum_{i=1}^I \xi_{ij}(1 - X_{ij}) + 2, \sum_{i=1}^I \xi_{ij}X_{ij} + 2, a = 0, b = 0.5)$

These make sense intuitively: estimates of  $p_k$  depend on the proportion of students who are estimated to have KC  $k$ ; estimates of  $g_j$  depend only on the performance of students who are missing one or more KC's for question  $j$ ; and estimates of  $s_j$  depend only on the performance of students who are estimated to have all the KC's needed for question  $j$ .

### 3.3 Estimation

Estimation is based on Markov chain Monte Carlo (MCMC) with Gibbs sampling (Patz & Junker, 1999; Junker & Sijtsma, 2001; Gelman et al., 2004; Wasserman, 2004). Our work is similar in spirit to Pardos et al. (2006), who used a maximum likelihood approach, but we go further in several ways. First, we estimate the population distribution of KC's, represented by the  $p_k$  parameters; second, we let the guess and slip probabilities vary across questions; and third, we include a constraint that the guess and slip probabilities must be between zero and one-half.

For the first step for each parameter's Markov chain each parameter is drawn from a Unif(0, 1) probability distribution, with the exception of  $\xi_{ij}$  which is set to 1 for all students and items. It is recommended in future analyses to begin chain at an estimate of the posterior distribution for parameter or to begin chain with maximum likelihood estimates (Gelman et al., 2004).

We implemented this Gibbs sampler in the R statistical package (e.g., Hornik, 2006). One run of 1000 steps requires approximately 11 hours on a DELL Xeon Precision 670 linux workstation (clock speed 3.6 GHZ). In general the first 500 steps were treated as burn in and we retained all other steps for posterior inferences. Visual inspection of time series plots of the MCMC output indicated good convergence to the stationary distribution.

## 4 Results

### 4.1 Item Prediction

To evaluate the accuracy of estimation of KC's for individual student we conducted cross-validation experiments with each of the six monthly data sets (October 2004 through March 2005) available for analysis, after imputing correct responses to not-reached scaffold questions (see Section 2).

We divided each month's data into a *training data set* and a *test data set*, each consisting of random halves of that month's main questions and scaffolds. Students with no data in either the training or test data set were omitted from both data sets for that month. Overall, student sample sizes for each of these 6 pairs of data sets ranged from a maximum around 650 each in October to a minimum around 435 each in March. The number of KC's assessed ranged from a minimum of 56 in October to a maximum of 77 in March. The number of questions (main and scaffold) ranged from a minimum of about 240 in October to a maximum around 630 in March.

We estimated the DINA model as described in Section 3 using the training data set to obtain per-student MAP (mode a-posteriori) estimates for each KC:  $\hat{\alpha}_{ik} = 1$  iff  $\hat{P}(\alpha_{ik} = 1 \mid \text{the data}) > 0.5$  as estimated from the MCMC procedure. Responses to questions in the test data set were then predicted using the rule

$$\hat{X}_{ij} = 1 \quad \text{iff} \quad \prod_{k=1}^K \hat{\alpha}_{ik}^{Q_{jk}} = 1$$

Note that this prediction assumes that  $s_j = g_j = 0$  for all questions  $j$  in the test data set. In our cross-validation scheme there is no opportunity to pre-calibrate questions in the training data set, so this seemed like a reasonable assumption. Finally, prediction accuracy for each student  $i$  is simply calculated as normalized complementary Hamming distance,

$$\frac{1}{T_i} \sum_{i,j \text{ in training set}} (1 - |\hat{X}_{ij} - X_{ij}|),$$

where  $T_i$  is the number responses from student  $i$  in the test data set.

Figure 3 illustrates the result of this procedure for the October, November and December data sets (top, middle and bottom panels). In each panel, the histogram on the left shows the distribution of prediction accuracies for all students in the training and testing data sets. In the scatter plot on

the right, each student is plotted according to the number of responses in the training data set for that student ( $X$ -axis) and the prediction accuracy for that student in the test data set ( $Y$ -axis). Thus, students with a low number of training responses have prediction accuracies fully ranging from 0% to 100%, whereas students with a high number of training responses have more stable prediction accuracies above 50%. The colors in these scatter plots indicate the number of responses for each student in the testing data set, to give a sense of the uncertainty in the estimated prediction accuracies. Figure 4 provides similar information for the three spring months, January, February and March.

In the Fall the average accuracy of prediction for questions held out for cross validation is relatively good: 69%, 72%, and 73% correct predictions for October, November and December, respectively (Figure 3). In the Spring average prediction accuracy for questions held out was slightly lower: 69%, 68%, and 70% for January, February, and March respectively (Figure 4). Anozie & Junker (2006) noted a similar but more severe dip in accuracy for predicting MCAS scores from ASSISTments around February in the same data set; they conjectured that this dip was due to a different experimental “forced scaffolding” regimen that was tried out in that month in live ASSISTment sessions.

Greater prediction accuracy in the fall may be due to the fact that students possess fewer KC’s, so that the base rate of incorrect responses is higher: if the model also tends to predict incorrect responses it will tend to be right. The base rates of incorrect responses decrease toward 50% as the school year progresses, making it much harder to accurately predict responses unless  $s_j$  and  $g_j$  are truly near zero<sup>6</sup>. Prediction accuracy might be improved by pre-calibrating  $s_j$  and  $g_j$  parameters for the test data set, rather than assuming they were all zero. Improvements in the transfer model, and greater focus of scaffold questions on measurement rather than tutoring, may also help.

Nevertheless, this is already a potentially powerful tool for teachers. Prediction accuracy of around 70% is high enough that reporting the underlying KC’s that students do and don’t know

---

<sup>6</sup>If by the end of the year, the base rate of incorrect response decreased to near zero, then the model could be accurate again by predicting the base rate response. However, student do not appear to master more than about 50-60% of the KC’s in this domain by year’s end (see Figure 5).

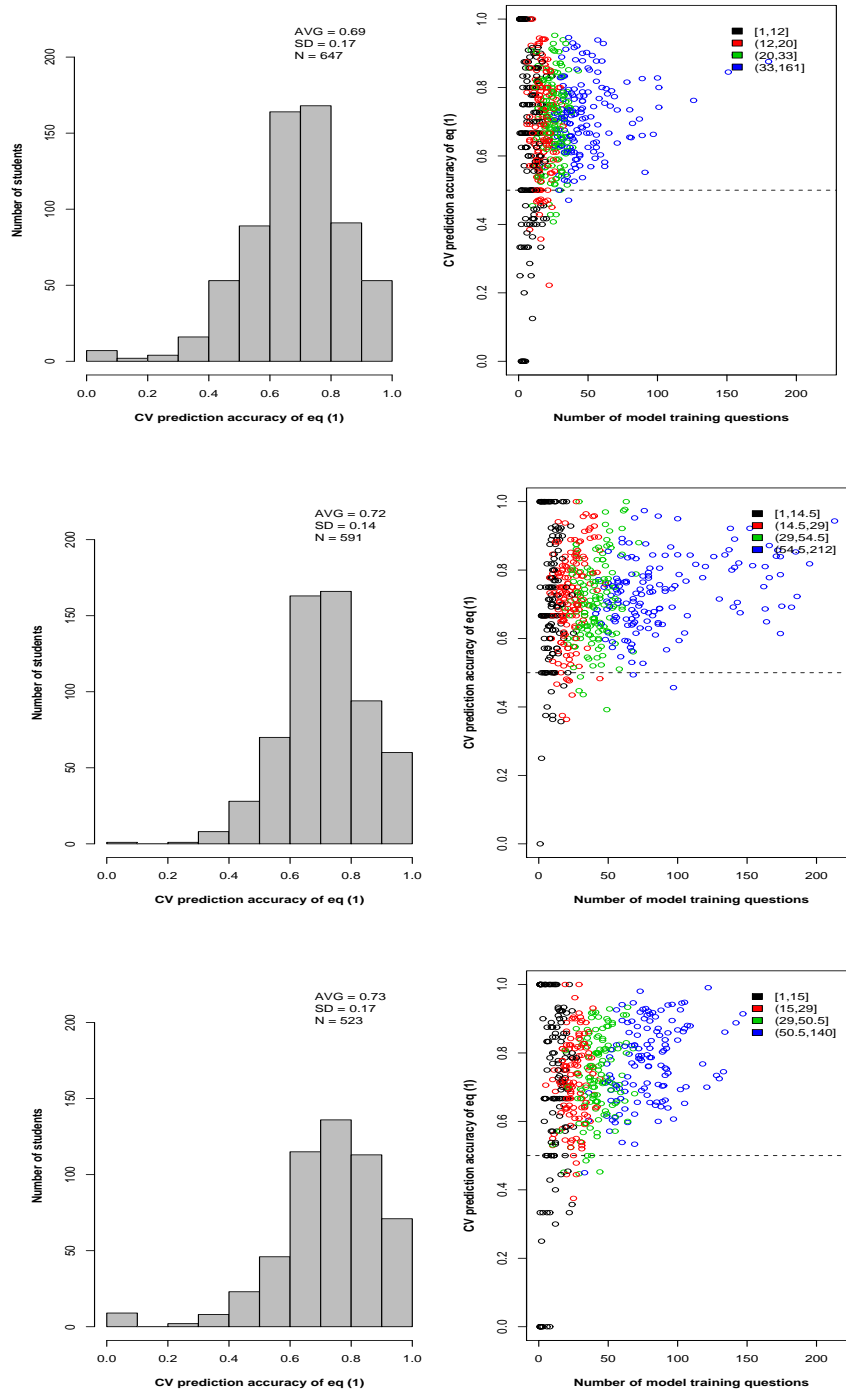


Figure 3: Split-half prediction accuracy, October (top), November (middle) and December (bottom). The histograms on the left show the distribution of prediction accuracies for all students in the training and testing data sets. The scatter plots on the right plot number of responses in the training data set vs. prediction accuracy for that student. The colors in these scatter plots indicate the number of responses for each student in the testing data set, to give a sense of the uncertainty in the estimated prediction accuracies.

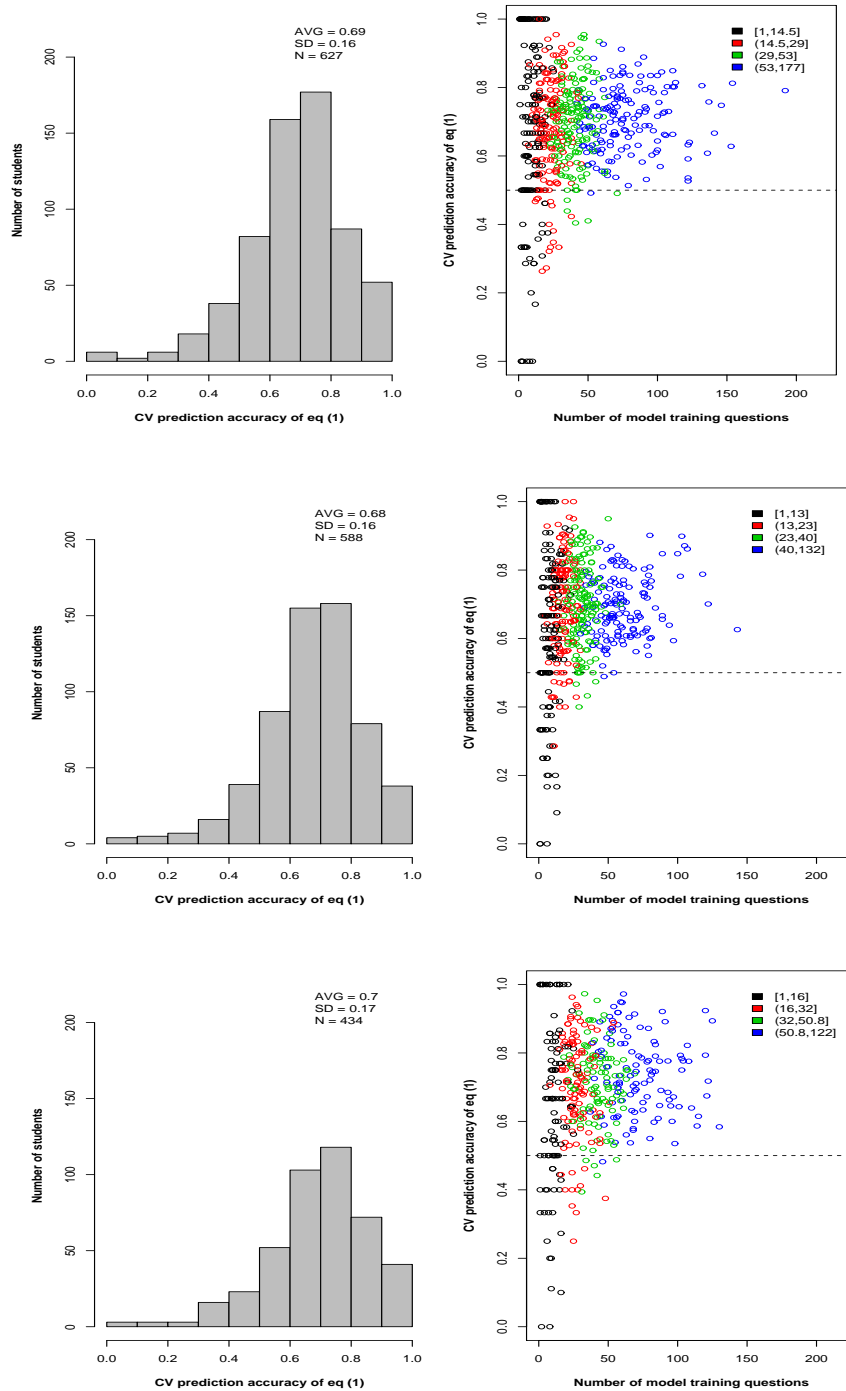


Figure 4: Split-half prediction accuracy, January (top), February (middle) and March (bottom). The histograms on the left show the distribution of prediction accuracies for all students in the training and testing data sets. The scatter plots on the right plot number of responses in the training data set vs. prediction accuracy for that student. The colors in these scatter plots indicate the number of responses for each student in the testing data set, to give a sense of the uncertainty in the estimated prediction accuracies.

would probably be informative for teachers' lesson plans, plans for review topics, etc., combined with teachers' prior knowledge of students.

## 4.2 Tracking student learning over time – Overall

We can also examine the proportion of KC's learned by each student in each month. For this, we used *all* of the post-imputation data available in each month, omitting students from each monthly data set who had no response data at all in that month.

We estimated the DINA model as before, on these larger monthly data sets, and from the estimated model we computed posterior 95% credible intervals for the probabilities  $P[\alpha_{ik} = 1 \mid \text{the data}]$ . We declared student  $i$  to have mastered KC  $k$ , if *both*  $\hat{P}[\alpha_{ik} = 1 \mid \text{the data}] > 0.5$ , *and* the equal-tailed 95% posterior interval for this probability did not contain 0.5.

Figure 5 shows the overall average number of KC's mastered by all students, using this criterion, for each month. There is a steady rise in learning over the course of the fall months, a small slump after the winter break, a precipitous drop in February as also noticed by Anozie & Junker (2006), and a recovery in March. At no time is the average proportion of KC's learned above 45%.

Figure 6 breaks the data in Figure 5 down into proportion of KC's mastered by each individual student. Students are ordered along the  $X$ -axis in each panel of Figure 6 in increasing order of KC's mastered in October. Aside from the overall trends also visible in Figure 5, we can see that, as the school year progresses, (a) KC mastery generally increases for most students (except for the February drop), and (b) students become more similar to one another, with respect to proportion of KC's learned.

## 4.3 Tracking student learning over time – By Knowledge Component

Figure 7 breaks Figure 5 down by KC. Abbreviated names for the KC's are labeled across the  $X$ -axis; for clarity the full names of the KC's are also listed in Table 1. Only 58 KC's could be compared across all six months of data. Above each KC are six colored dots labeled with the month letter: (O)ctober, (N)ovember, (D)ecember, (J)anuary, (F)ebruary, or (M)arch, indicating

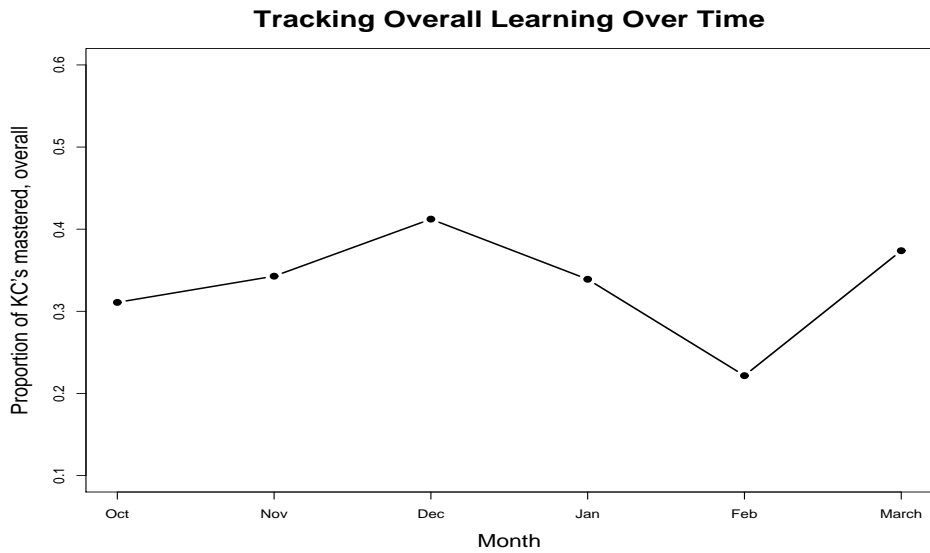


Figure 5: Proportion of KC's mastered, averaged over all students in each month (see text for definition of mastery).

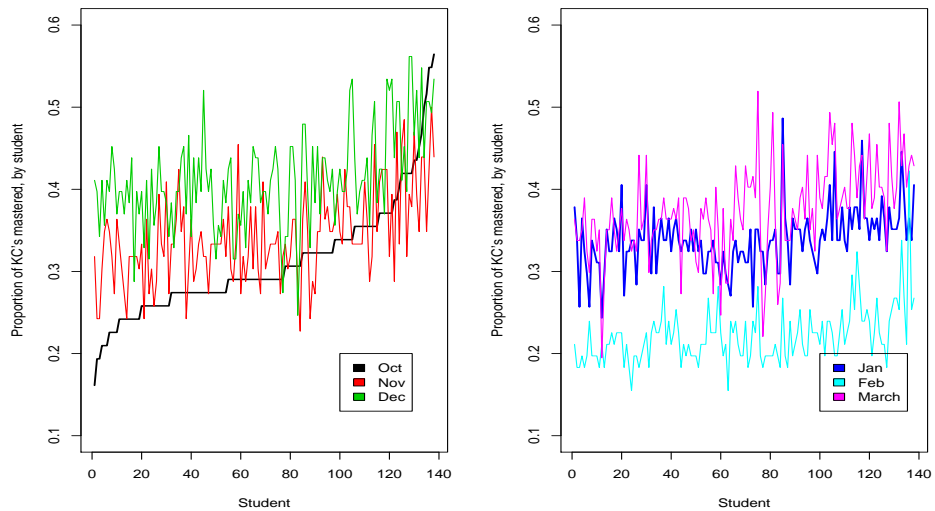


Figure 6: Proportion of KC's mastered per student (see text for definition of mastery). Different colored curves correspond to different months.

### Tracking student Learning Over time– By Knowledge Component

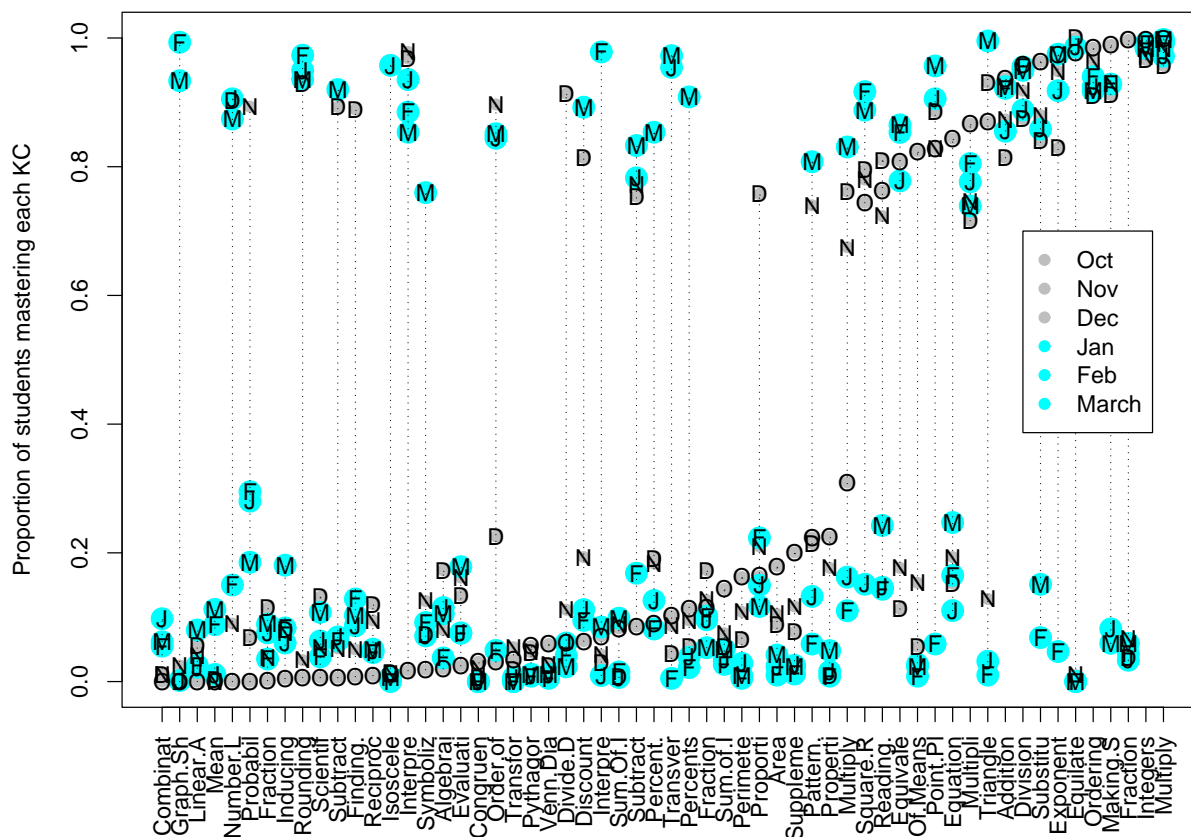


Figure 7: Proportion of students mastering each KC in each month (see text for definition of mastery). Full names of KC's are listed in Table 1. KC's are listed in increasing order of mastery as of October 2004. Colored dots indicate proportion of students who mastered each skill in each month (months indicated by letter in dot).

the proportion of students we estimate to have mastered that KC in each month.

There is evidence of both learning and forgetting in Figure 7. KC's that are easy throughout the year include Addition, Integers, and Multiplying Positive and Negative Numbers. KC's that are difficult throughout the year include Combinatorics, Linear/Area/Volume Conversion, and Venn Diagrams. KC's on which students generally improved throughout the year include Rounding, Subtracting Decimals, and Finding Percents. KC's on which students seemed to fall down by the end of the year included Making Sense of Expressions and Equations, and Fractions.

Despite the ambiguous picture of learning in Figure 7, there is in fact evidence that students

	KC Definition		KC Definition
1	Combinatorics	30	Transversals
2	Graph.Shape	31	Percents
3	Linear.Area.Volume.Conversion	32	Fraction.Multiplication
4	Mean	33	Sum.of.Interior.Angles.Triangle
5	Number.Line	34	Perimeter
6	Probability	35	Proportion
7	Fraction.Division	36	Area
8	Inducing.Functions	37	Supplementary.Angles
9	Rounding	38	Pattern.Finding
10	Scientific.Notation	39	Properties.of.Geometric.Figures
11	Subtracting.Decimals	40	Multiplying.Decimals
12	Finding.Percents	41	Square.Root
13	Reciprocal	42	Reading.graph
14	Isosceles.Triangle	43	Equivalent.Fractions.Decimals.Percents
15	Interpreting.Numberline	44	Of.Means.Multiply
16	Symbolization.Articulation	45	Point.Plotting
17	Algebraic.Manipulation	46	Equation.Solving
18	Evaluating.Functions	47	Multiplication
19	Congruence	48	Triangle.Inequality
20	Order.of.Operations	49	Addition
21	Transformations.Rotations	50	Division
22	Pythagorean.theorem	51	Substitution
23	Venn.Diagram	52	Exponents
24	Divide.Decimals	53	Equilateral.Triangle
25	Discount	54	Ordering.Numbers
26	Interpreting.Linear.Equations	55	Making.Sense.of.Expressions.and.Equations
27	Sum.Of.Interior.Angles.more.than.3.Sides	56	Fractions
28	Subtraction	57	Integers
29	Percent.Of	58	Multiplying.Positive.Negative.Numbers

Table 1: Full names of the knowledge components (KC's) listed in Figure 7.

are learning, in part due to normal classroom experiences, and in part due to the ASSISTments tutor itself (Razzaq, et al., 2005). Some of the mastery estimates shown in Figure 7 are based on relatively small student and item sample sizes within a particular month, and therefore have not moved the posterior probability of possessing the KC appreciably from the prior value of 0.30, which does not meet our criterion of mastery. However, our analysis does not yet take into account the fact that in previous months students may have shown mastery of a KC. We are currently investigating two approaches to incorporating this information: one is to use the previous month's proportion mastered as the prior probability of possessing each KC when estimating the DINA model in the current month; another is to model each KC indicator  $\alpha_{ik}$  as a two-state Markov chain with a transient “unlearned” state and an absorbing “learned” state, by analogy with the knowledge tracing algorithm of Corbett, Anderson & O'Brien (1995).

#### **4.4 Evaluating the quality of questions and transfer model**

Our analysis is also informative about the nature of the transfer model and its use in tagging questions with KC's.

Consider, for example, Figure 8, which shows a main ASSISTment question and its associated scaffold questions. Figure 9 shows posterior boxplots for the slip parameters for the main question and each scaffold question, estimated from each of our six monthly data sets. Figure 10 gives the analogous plots for the guessing parameters for these same questions.

To the extent that the scaffolding questions have lower slip and guess parameters than the main question, they are more reliable indicators of the KC than the main question is. DiBello, Stout and Roussos (1995) refer to this increased per-item reliability in measuring KC's as “high positivity” for the transfer model.

However, another phenomenon appears in Figures 9 and 10 as well: in most months, the slip parameter decreases, and the guessing parameter increases, from one scaffold question to the next. These trends tell us that the scaffolds are getting successively easier, perhaps reflecting the fact that the student does not have to re-parse the problem set-up once he/she has parsed it for the main question (and perhaps the first scaffold), and/or a practice effect with the KC. This reflects a validity

A.

C.

Table 1: Which graph above contains the points in the table below?

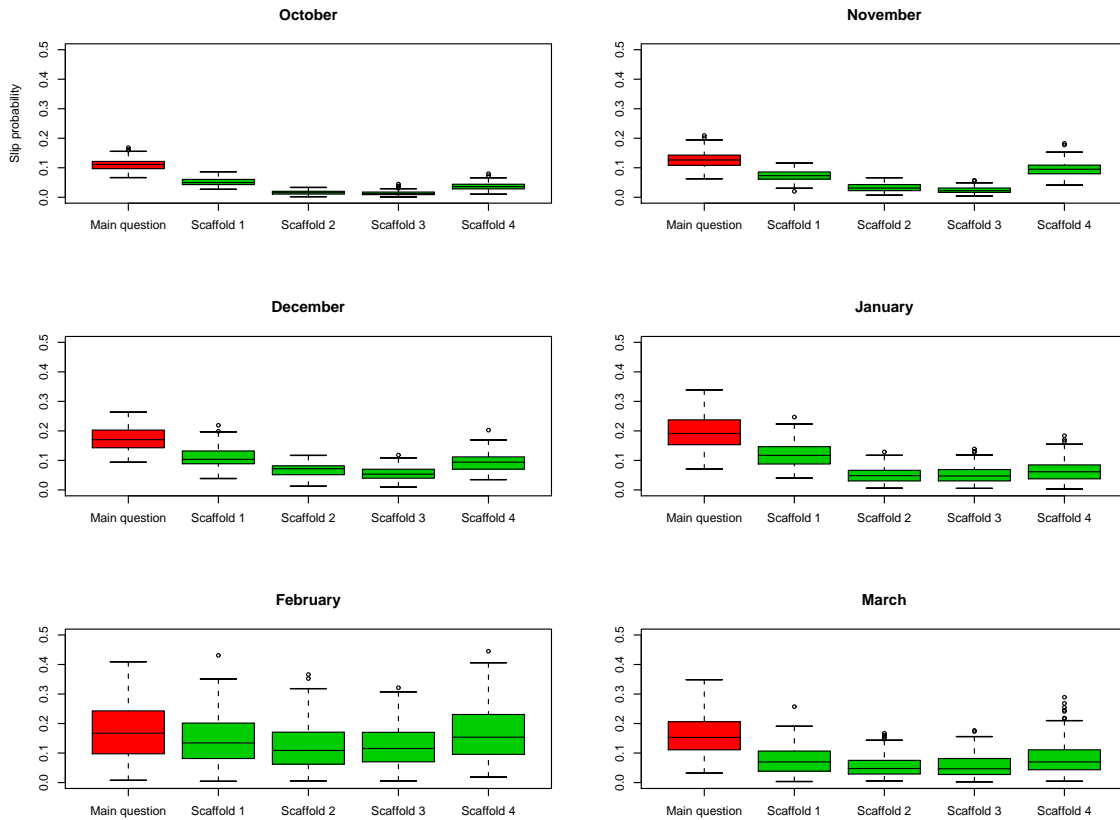
$x$	$y$
-2	-3
-1	-1
1	3

B.

D.

1. Let's go through the steps for solving the problem. The first point in the table is  $(-2, -3)$ . The  $x$ -coordinate is negative and the  $y$ -coordinate is negative. *If you plotted this point, which quadrant would it be in?*
2. Good. The next point is  $(-1, -1)$ . *Which quadrant would this be plotted in?*
3. Okay. The final point is  $(1, 3)$ . *Which quadrant is this point in?*
4. Great. You know that the correct graph has a line that passes through the first and third quadrants. That means you can eliminate some choices. Based on the location of the plotted points, *which graph do you think is the correct one?*

Figure 8: Illustration of an ASSISTment main question (top half) with its associated scaffold questions (bottom half).



Scaffold 1: Let's go through the steps for solving the problem. The first point in the table is  $(-2, -3)$ . The  $x$ -coordinate is negative and the  $y$ -coordinate is negative. If you plotted this point, which quadrant would it be in?

Scaffold 2: Good. The next point is  $(-1, -1)$ . Which quadrant would this be plotted in?

Scaffold 3: The final point is  $(1, 3)$ . Which quadrant is this point in?

Scaffold 4: Great. You know that the correct graph has a line that passes through the first and third quadrants. That means you can eliminate some choices. Based on the location of the plotted points, which graph do you think is the correct one?

Figure 9: Posterior boxplots for slip parameter estimates, for the main question and scaffolding questions for Figure 8. The line in the middle of the box represents the posterior median, the edges of the box the posterior 25<sup>th</sup> and 75<sup>th</sup> percentiles, and so forth.

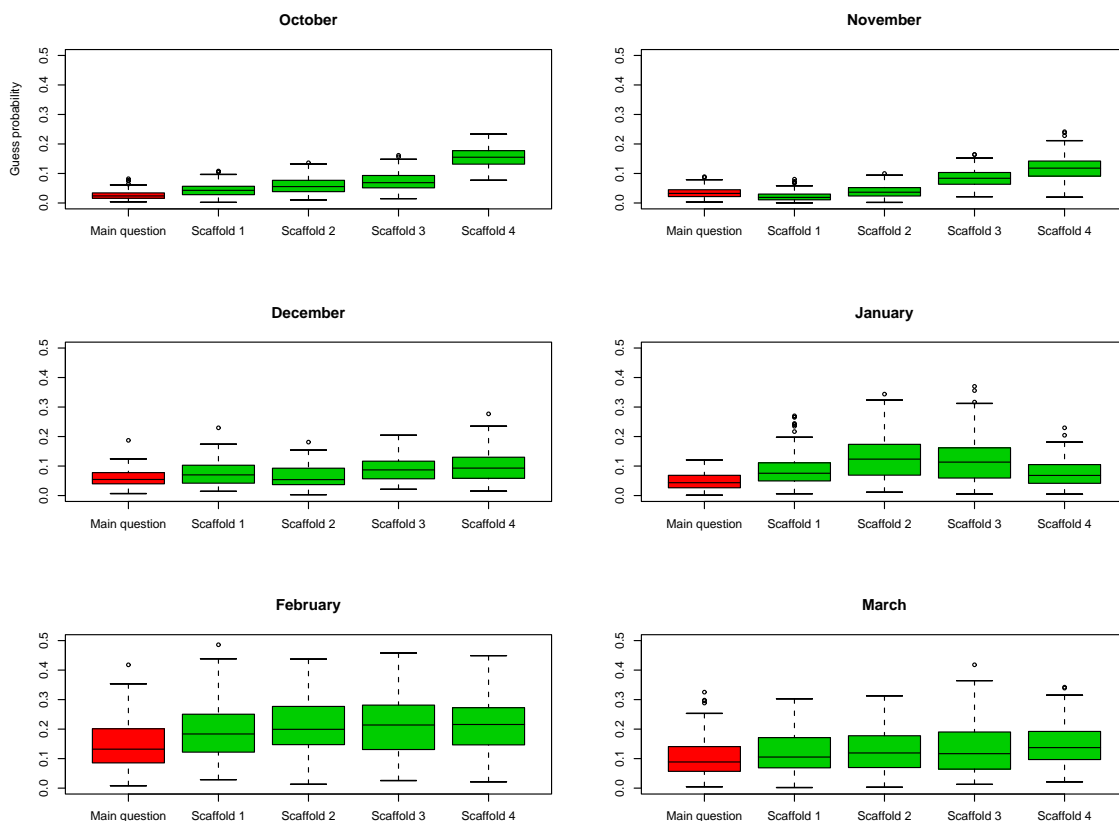


Figure 10: Posterior boxplots for guessing parameter estimates, for the main question and scaffolding questions for Figure 8. The line in the middle of the box represents the posterior median, the edges of the box the posterior 25<sup>th</sup> and 75<sup>th</sup> percentiles, and so forth.

decision about the “completeness”, to use DiBello et al.’s (1995) term, of the transfer model: there is a tradeoff to make between developing a more complete list of KC’s and other determinants of student performance (reducing biases in assessing whether KC’s have been learned or not), vs. having little unique information about each individual component of the model (increasing uncertainty about whether KC’s have been learned or not).

## 5 Discussion

The ASSISTment System was conceived and designed to help classroom teachers address the accountability demands of the public education system in two ways. First, ASSISTments provide

ongoing benchmarking of students that can be used to predict success on end-of-year accountability exams, while providing some instructional benefit—not all time spent with ASSISTments is lost to testing. Second, the system can provide feedback to teachers on students’ progress in specific areas or on specific sets of KC’s. Anecdotal evidence suggests that teachers are positive about the system, and students are impressed with its ability to track their work. Previous work, e.g. as surveyed by Junker (2006), has concentrated on the statistical problem of predicting end-of-year MCAS exam scores from ASSISTments data.

The present paper turns to the other task, that of diagnosing and reporting achievement of students learning particular knowledge components (KC’s) in a transfer model. Figure 11 shows a knowledge components report for teachers, based on crediting/blaming the most difficult KC involved in each correct/incorrect ASSISTment question (similar to Feng, Heffernan, Mani & Heffernan’s, 2006, max-difficulty reduction of the transfer model). This approach clearly under-credits less-difficult KC’s involved in multiple-KC items. Use of the DINA model, as illustrated in this paper, has the potential to more accurately diagnose whether students do or do not possess particular KC’s in a transfer model.

As such, we plan to turn to the problem of developing reports like Figure 11 from the DINA model. In order to do this, we will have to speed up estimation of KC’s; current MCMC methods using the DINA model are quite slow. On the other hand, perhaps some of the suggestions of Junker & Sijtsma (2001) for data summaries that are relevant about KC acquisition will be helpful to us.

A further refinement of the DINA approach, now under investigation, combines the knowledge tracing algorithm of Corbett, Anderson & O’Brien (1995) with Bayes Net (DINA) models (Junker & Sijtsma, 2001). This has the potential advantage of smoothing over artificial backsliding such as the February dip in the 2004–2005 data we considered for this paper.

Another aspect of the project is that the ASSISTment system must serve a variety of stakeholders, and not all of them need or want reports at the same level of granularity. Indeed, the ASSISTment project has worked with four different transfer models, from a one-variable Rasch model, which is likely best for predicting MCAS scores, to a 106-KC Bayes Net model, which may

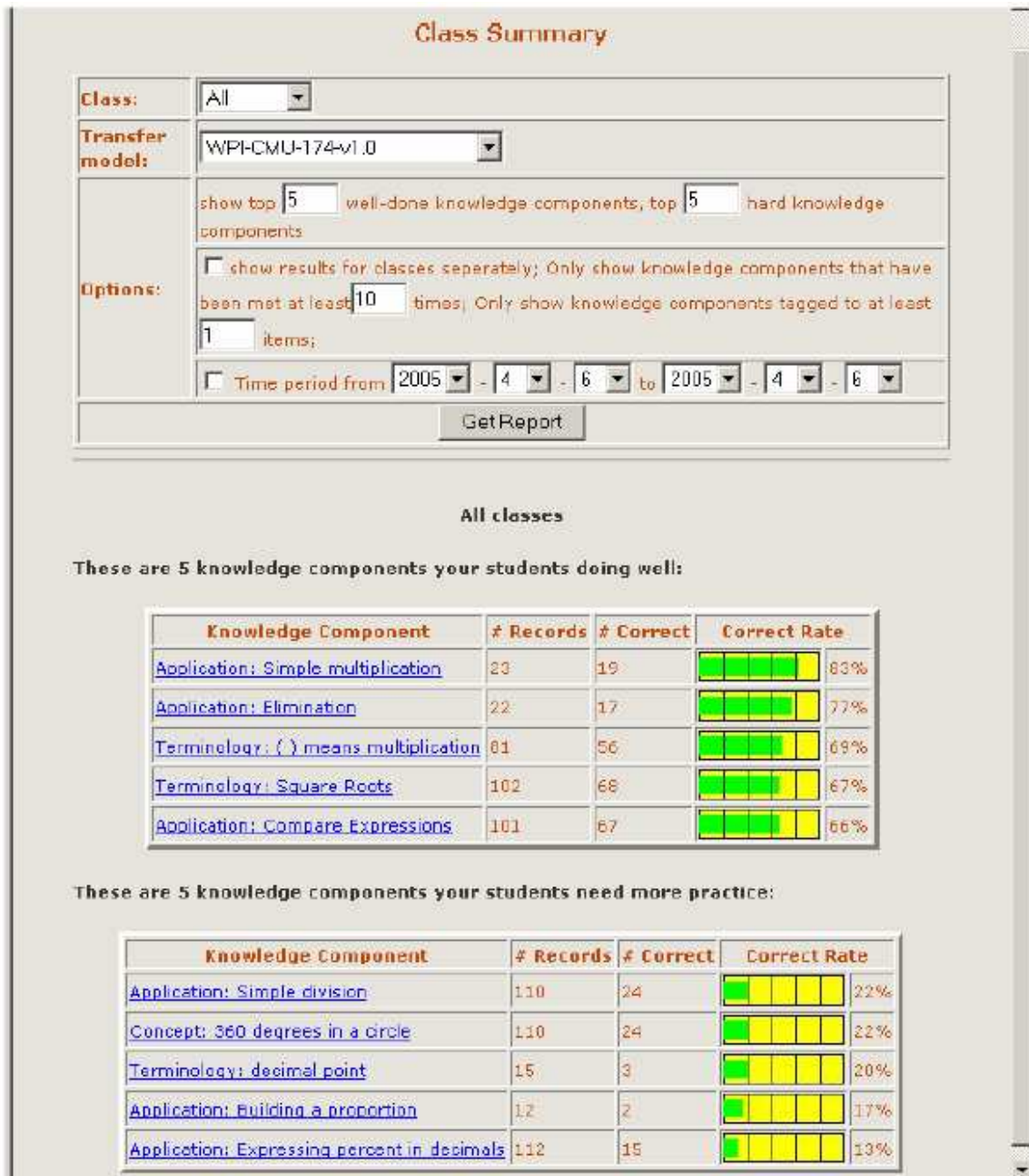


Figure 11: Classroom-level KC's report for teachers in the ASSISTment system. Student-level gradebook information is also available in the system.

be closer to optimal for providing teacher feedback. As the ASSISTment system is considered in multiple States and other jurisdictions, additional transfer models will be needed, that are aligned to those States' learning standards.

The multiple transfer-model problem becomes more acute when considering the information that scaffold questions provide for inferences about students. It may be possible to write scaffold questions that tap one KC at a time in a particular transfer model, but the same questions may tap more than one KC at a time in a finer-grained transfer model; or they may tap bits and pieces of KC's in a transfer model that is not a proper coarsening or refinement of the transfer model used to develop the scaffold questions. In addition, question developers sometimes write scaffolds based on KC-related goals, and sometimes based on tutorial goals, for example reframing part or all of a question to look at the same KC in a different way. This may make KC learning look less stable than it really is, since students' KC-related behavior is also influenced by the effectiveness of the tutorial reframing. In part to understand this, we are currently building some true one-KC questions to investigate the stability of KC's across questions.

## References

- Anozie, N.O. and Junker, B. W. (2006). *Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system*. American Association for Artificial Intelligence Workshop on Educational Data Mining (AAAI-06), July 17, 2006, Boston, MA.
- Ayers, E. and Junker, B. W. (2006). *Do skills combine additively to predict task difficulty in eighth-grade mathematics?* American Association for Artificial Intelligence Workshop on Educational Data Mining (AAAI-06), July 17, 2006, Boston, MA.
- Barnes, T. (2005). Q-matrix Method: Mining Student Response Data for Knowledge. *Proceedings of the AAAI-05 Workshop on Educational Data Mining, Pittsburgh, 2005 (AAAI Technical Report #WS-05-02)*.
- Beck, E.J, Peng, J, Mostow, J. (2003). Assessing Student Proficiency in a Reading Tutor that Listens. *Proceedings of the 9th International Conference on User Modeling, Johnstown, PA*. Preprint available at [http://www.cs.cmu.edu/~listen/pdfs/UM2003\\_paper\\_test\\_prediction.pdf](http://www.cs.cmu.edu/~listen/pdfs/UM2003_paper_test_prediction.pdf).
- Beck, J.E.; Jia, P. and Mostow, J. (2004). Automatically assessing oral reading fluency in a computer tutor that listens. *Technology, Instruction, Cognition and Learning 2*: 61-81.
- Corbett, A. T., Anderson, J. R. and O'Brien, A. T. (1995) Student modeling in the ACT programming

- tutor. Chapter 2 in P. Nichols, S. Chipman, & R. Brennan, eds., *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Erlbaum.
- Corbett, A. T., Koedinger, K. R. and Hadley, W. H. (2001). Cognitive Tutors: From the research classroom to all classrooms. In Goodman, P. S. (Ed.), *Technology Enhanced Learning: Opportunities for Change*. Mahwah, NJ: Lawrence Erlbaum Associates.
- DiBello, L. V., Stout, W. F. and Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. Chapter 15 in Nichols, P. D., Chipman, S. F. and Brennan, R. L. (eds.) (1995). *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Embretson, S. E. (1984). A General Latent Trait Model for Response Processes. *Psychometrika*, 49, 175–186.
- Embretson, S. E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika*, 64, 407–433.
- Feng, M., Heffernan, N., Mani, M., and Heffernan, C. (2006). *Using mixed effects modeling to compare different grain-sized skill models*. American Association for Artificial Intelligence Workshop on Educational Data Mining (AAAI-06), July 17, 2006, Boston, MA.
- Gelman, A.; Carlin, J. (2004). *Bayesian Data Analysis: Second Edition*. Boca Raton, Fl.: Chapman & Hall / CRC.
- Haertel, H.E. (1989). Using Restricted Latent Class Models to Map Structure of Achievement Items. *Journal of Educational Measurement* 26: 301-321.
- Hamilton, L. (1992). *Regression With Graphics: A Second Course in Applied Statistics*. Belmont, California.: Duxbury Press,
- Heffernan, N.T., Koedinger, K.R. and Junker, B.W. (2001). Using Web-Based Cognitive Assessment Systems for Predicting Student Performance on State Exams. Technical Report, Institute of Educational Statistics: US Dept. of Education. Dept. of Computer Science, Worcester Polytechnic Institute Univ.
- Hornik, K. (2006). *The R FAQ*. ISBN 3-900051-08-9, available online at <http://CRAN.R-project.org/doc/FAQ/R-FAQ.html>.
- Junker, B.W. (1999). Some statistical models and computational methods that may be useful for cognitively-relevant assessment. Prepared for the National Research Council Committee on the Foundations of Assessment. Available online at <http://www.stat.cmu.edu/~brian/nrc/cfa/>.
- Junker, B.W. (2006). 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 setting*. Maple Grove, MN: JAM Press. Preprint available at <http://www.assistment.org/project/>.
- Junker, B.W. and Sijtsma K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement* 25: 258-272.

- Koedinger, K. R.; Anderson, J. R.; Hadley, W. H. and Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education* 8:30-43.
- Koedinger, K. R.; Corbett, A. T.; Ritter, S.; and Shapiro, L. J. (2000). Carnegie Learning's Cognitive Tutor: Summary research results. White Paper. Technical Report, Carnegie Learning, Pittsburgh, PA.
- Macready, G. B. and Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, 2, 99–120.
- Mislevy, R.J. and Wu, R-K. (1996). Missing responses and Bayesian IRT ability estimation: Omits, choice, time limits, and adaptive testing (Research Report RR-96-30-ONR). Princeton, NJ: Educational Testing Service.
- Olson, L. (2005). State Test Programs Mushroom as NCLB Mandate Kicks In. *Education Week*, Nov. 30: 10-14.
- Pardos, Z. A.; Heffernan, N. T.; Anderson, B. and Heffernan, C. L. (2006). Using fine-grained skill models to fit student performance with Bayesian networks. Submitted to the International Conference on Intelligent Tutoring Systems (ITS 2006): Education Data Mining Workshop, Jhongli, Taiwan.
- Patz, R. J. and Junker B. W. (1999). Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses. *Journal of Educational and Behavioral Statistics* 24: 342-366.
- Razzaq, L.; Feng, M.; Nuzzo-Jones, G.; Heffernan, N.T.; Koedinger, K. R.; Junker, B.; Ritter, S.; Knight, A.; Aniszczyk, C.; Choksey, S.; Livak, T.; Mercado, E.; Turner, T.E.; Upalekar. R, Walonosk.; J.A., Macasek. M.A.; Rasmussen, K.P. (2005). The Assistentment Project: Blending Assessment and Assisting. In Looi, C.K.; McCalla, G.; Bredeweg, B. and Breuker, J. (Eds.) Proceedings of the 12th International Conference on Artificial Intelligence In Education, 555-562. Amsterdam: ISO Press.
- Tatsuoka, K.K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, and M.G. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 453-488). Hillsdale, NJ: Erlbaum.
- de la Torre, J., and Douglas J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika* 69: 333-353.
- Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. New York, NY.: Springer-Verlag.

Websites:

<http://www.assistentment.org>

<http://www.learnlab.org>

<http://www.educationaldatamining.org>